Introduction

CTB/McGraw-Hill presents a unified, integrated response to *SMARTER Balanced RFP No. 14, SBAC Pilot Item/Task/Stimulus Research, Development, and Reviews.* CTB has partnered with five other companies to create a robust Collaborative that will provide the breadth and depth of expertise needed to meet the ground breaking requirements this SBAC proposal. Our Collaborative includes:

- CTB/McGraw-Hill (CTB)
- American Institutes for Research (AIR)
- Data Recognition Corporation (DRC)
- Council for Aid to Education (CAE)
- College Board (CB) and
- Human Resources Research Organization (HumRRO).

Our combination of organizational expertise and experience as well as our current engagement in innovative assessment practices and technology practices provide a solid foundation for the development of the final SBAC balanced assessment system. All of our members agree that working with educators to develop large scale assessments will bring incredible content expertise and understanding of how students interact with the content and when combined with our industry background/resources, will provide powerful products that have instructional sensitivity. Additionally, we recognize and value the need to work collaboratively with other SBAC contracts to assure the delivery of the next generation of assessments to educators across the county.

Each organization brings to this endeavor unique and complementary expertise and capacity that are crucial to the development and evaluation of a final pool of stimuli, items, and performances for the 2012–2013 Pilot Test.

CTB/McGraw-Hill has a proven history of successfully designing, developing and delivering summative and interim assessments. CTB has experience in all areas of large-scale assessment. Our particular expertise in the following key areas is directly related to the requirements of this RFP:

- developing items that align to the Common Core State Standards (CCSS)
- employing evidence-based assessment development procedures
- conducting content, bias/sensitivity, and accessibility reviews
- writing items and performance tasks using teacher authors
- creating research designs for adaptive testing instruments

CTB worked in partnership with AIR to deliver the first state-wide, adaptive norm-referenced *TerraNova* assessment. CTB works exceptionally well with the partners in the Collaborative and understands and respects that the Collaborative must work with SBAC leadership, state members, and other partners.

AIR has demonstrated the ability to be the SBAC vendor-partner that reaches beyond state-of-the-art practices and delivers innovative models, materials, and strategies that advance the field of measurement and improve the validity of item and task scores. In student assessment, their commitment is evidenced by success in delivering the following:

- the first and only NCLB-approved, full-year, multiple-opportunity, online adaptive assessments, now operational in four states
- more than 4,000 operational, innovative, constructed-response items scored by computer in real-time; item sharing of summative multiple-choice and technology-enhanced items across states
- with CTB/McGraw Hill, the first state-wide, adaptive norm-referenced assessments (*TerraNova*)
- the first statewide, online adaptive formative assessments delivered in the same system as summative assessments and measuring the same constructs on the same scale.

DRC brings significant item development expertise to the Collaborative, including current experience working directly with teachers developing items for the Michigan assessment program. DRC's Test Development Team has extensive knowledge of the Common Core State Standards (CCSS), and has experience working with several states in developing items and aligning state standards to the CCSS, including Alabama, Alaska, Louisiana, and Pennsylvania. DRC has a long and successful history of working collaboratively with other contractors and clients, focusing on effective communication and mutual respect.

CAE has a long-standing record of creating and administering open-ended performance assessments for the secondary and higher education sectors. As a result of this, the Gates Foundation and the Organization for Economic Co-operation and Development (OECD) independently engaged CAE to develop new national and international performance tasks. CAE has the capacity to form a large item-development team within a short time frame and deliver innovative performance tasks for the SBAC Pilot Test Item Pool.

The College Board has a goal of ensuring that every student has the opportunity to prepare to become college ready as well as to enroll in and graduate from college. Their College Readiness initiatives promote curricula, assessment tools, and guidance resources that help K-12 students prepare for the academic rigors of higher education. College Board championed innovation, equity and excellence for generations of students. They work to empower teachers by training teachers to become skilled, dedicated, and inspiring to the students they teach in a variety of content areas. College Board will bring this expertise to the development of items to meet the SBAC goal of college- and career-readiness (CCR) for students.

HumRRO conducts outstanding research, program evaluation, and policy analysis. The company has experience in working with and supporting a wide variety of state agencies, federal agencies, foundations, and corporate/nonprofit organizations. HumRRO is nationally recognized and respected for conducting the special studies in the educational arena. HumRRO currently serves as independent evaluator for the National Assessment of Educational Progress (NAEP).

CTB and all of the highly motivated member companies are excited to be part of the Collaborative. More importantly, we believe we can complete all of the pioneering work set forth in this proposal within the timelines. In order to complete the work, we have built strong, experienced program and project management teams across the Collaborative. They will manage the work of all of the companies across the four tasks described below while working closely with SBAC leadership and member states. While all organizations within our collaborative contribute to the planning and execution of each task, the primary responsibilities of each organization are charted in the following Tables 1-4:

Table 1: Part 1: Oversight, Item/Task/Stimulus Development and Reviews

Activity	Lead Organization	Participants
Oversight	CTB	All
Conventional item development and reviews	CTB	AIR, DRC
Technology-enhanced item development and reviews	CTB	AIR, DRC
Performance task development and reviews	CTB	CAE
CCR reviews	College Board	All

Table 2: Automated Scoring and Scoring Models

Activity	Lead Organization	Participants
Automated scoring research	AIR	DRC

Table 3: Item/Task/Stimulus Research and Development

Activity	Lead Organization	Participants
Research: Cognitive labs and small-scale trails	AIR	CTB, DRC
Development of TE items	AIR	CTB, DRC

Table 4: Study of Item Procurement Options

Activity	Lead Organization	Participants
Study design and implementation	HumRRO	CTB
Study evaluation	HumRRO	

The tables above outline the extensive experience and expertise we would bring to the Consortium, but the enclosed proposal offers all of the thoughtful details. Building on our solid foundation of assessment expertise, research, and large scale experience, we offer the following innovations to the development of the Pilot Test item pool:

- Pushing the industry in the development of cutting edge technology type items and complex performance tasks to include innovative stand-alone technology-enhanced items
- A unique approach to working with teacher authors to focus on claims and evidence
- Open-source technology for administration of items for cog-labs and small-scale trials
- A cost-effective plan for online reviews of stimuli, items, and tasks

One of the major strengths of this Collaborative are the expert, talented staff and large network of resources that we bring to the development and delivery of the stimuli/texts and items and performance tasks for the cognitive labs and small-scale trials and other tasks. This Collaborative can leverage the capacity and resources of each company to support our solutions as defined in each of the four tasks enclosed in our technical proposal. Again, our Collaborative is eager to present our expertise and innovative solutions to SBAC as per the requirements of this RFP. The details of the CTB, DRC, AIR, Council for Aid to Education (CAE), College Board and HumRRO technical approach to these innovations and other requirements begin in the next section.

Technical Proposal

OBJECTIVE AND SCOPE OF WORK

1. Deliver sufficient items/tasks, aligned with item/task specifications and content specifications, for 2012-2013 pilot testing for purposes such as scaling and linking studies, multidimensionality studies, and validation of automated scoring

The Collaborative has been formed to bring together the strengths, capacity, and resources of each company—CTB, DRC, AIR, the College Board, and the Council for Aid to Education (CAE)—to meet the challenges of the scope of work for each task and deliver quality assessment items for the cognitive labs, small-scale trials, and the 2012-2013 pilot testing. Given the large number of items to be developed in a very short period of time, in addition to involving many participants from SBAC member states, we will mobilize the Collaborative's staff and extensive network of talented resources to meet the schedule. Our resources will develop and deliver the stimuli/texts and items and tasks for the cognitive labs and small-scale trials in spring 2012. They will also contribute to the development of the more than 10,000 items and tasks for pilot testing during the 2012–2013 school year.

The Collaborative partners have extensive experience and expertise in working with educators to develop assessments. For example, CTB and DRC have worked extensively with educators from many states and countries—including Arizona, Connecticut, Maryland, Missouri, Nebraska, North Dakota, Pennsylvania, Bermuda, and Qatar—to train them in the art and science of writing items and tasks. In addition to content expertise, teachers bring to the item development or review task a deep understanding of how students interact with the content and demonstrate understanding—or misconceptions—about the content. Typically, classroom teachers do not have experience working with item specifications or an evidence-centered design approach to assessment. However, the Collaborative's assessment specialists will work closely with participants to help them understand and apply the item/task/stimulus specifications or the content, bias/sensitivity, accessibility guidelines to produce the highest quality items/tasks/stimuli possible.

We will use the specifications developed by the SBAC 04 vendor and the review training materials and guidelines developed by the SBAC 08 vendor to train and coach participants in how to unpack and apply the specifications or guidelines. As we work with SBAC educators to develop assessments aligned to the Common Core State Standards, we will focus on using an evidence-centered approach. We will help educators integrate the specifications and guidelines with their knowledge and experience to more deeply understand the constructs being assessed and how assessment is the process of eliciting evidence that supports inferences about students' mastery of the content.

We are excited to present our innovative and effective plan for developing items aligned with the item/task specifications and content specifications in an efficient and expeditious approach that maintains quality, minimizes risk, and maximizes the professional development of educator participants. Our plan for how we will work with educator item/task writers is described in detail in section L1-12.

Our item development plan will build educators' assessment literacy—skills they will take back to their classrooms, schools, and districts. Our plan is based on forming small teams of three educator participants and one editor from the Collaborative partners. Each team will focus on developing items for a portion of the Common Core State Standards across two or three grade levels. By forming small cross-grade teams, participants will become intimately familiar with a set of standards and item specifications, and deepen their understanding of how the content represented by the set of standards and students' mastery of the content grow and change across the grade levels.

The team members will work closely together with their editor to plan, discuss, write, and review draft items produced by the team. As the team works on items for a given claim and standard, the editor will facilitate a review and discussion of the relevant item specifications with the team in greater depth than during the item/task writer training. The team will plan its overall approach to the items to write for the standard(s) based on the item development plan and specifications. As the writers draft and submit items, they will receive focused feedback from their editor. This will further the goal of developing educators' understanding of the content, the item specifications, and their assessment literacy skills. The team will come together regularly via webinars to review and discuss draft items and to resolve issues. Any given team will focus on just a portion of the content and item specifications. But the educators will develop such a deep internalized understanding of how to use the item specifications to develop high-quality items that their knowledge and skills should transfer easily to the broader content domain.

In the following sections we describe our proposed plan for hiring item/task writers and reviewers and training them. Our assumption for the number of writers or reviewers needed is based on our item development plan and the number of items or stimuli needed. Therefore, we present tables in this section to provide an overview of our proposed plan. However, please see section L1-12 for a more complete description of our proposed item development plan.

Cognitive Lab and Small-Scale Trials Item Development Plan

We understand the criticality of developing exemplary items for the cognitive labs and the small-scale trials, and the importance of exploring how to assess standards that have traditionally not been assessed using conventional item formats, how best to create and deliver technology-enhanced and technology-enabled items, and how to ensure accessibility of items to all students.

Given the need to produce items for the cognitive labs and small-scale trials in a very short period of time, the Collaborative proposes that all items/tasks/stimuli for both the cognitive labs and the small-scale trials be produced by the members of the Collaborative and our experienced item development resources. The development of these items provides the opportunity for the Collaborative's item writers to provide feedback on the clarity and use of the item specifications, so that SBAC may make refinements, if any, prior to using the specifications for the large-scale development of the 2012–2013 pilot items.

The table below presents our plan for the development of the items for the cognitive labs and small-scale trials. For the purpose of preparing our price proposal, we made the assumption that the number of items developed for each type would be in the same proportion as for the 2012–2013 pilot items. Because these are experimental items, we will not develop an overage of items in case of loss during editorial reviews or as a result of the labs or trials. (We understand, however, that some items may indeed become part of the pool of items for the 2012–2013 pilots.) These items will go through internal editorial reviews but will not be presented to committees of SBAC educators for content, bias and sensitivity, or accessibility reviews. Should the final item development plan differ significantly from our assumptions, the CTB program manager will initiate the change management process to make appropriate adjustments to the scope and price.

Table 5: Cognitive Lab and Small-Scale Trials Item Development Plan

Content Area	Item Type	Cognitive Lab Items	Small-Scale Trials Items
ELA	Selected Response	49	309
	Technology-Enhanced Selected Response	64	442
	Constructed Response	56	379
	Technology-Enhanced Constructed Response	70	470
	Performance Task	7	12
	Technology-Enhanced Performance Task	14	20
	Stimuli	30	180
Math	Selected Response	49	321
	Technology-Enhanced Selected Response	78	545
	Constructed Response	28	189
	Technology-Enhanced Constructed Response	84	545
	Performance Task	7	12
	Technology-Enhanced Performance Task	14	20
Total Items		520	3264

2012–2013 Pilot Item Development Plan

The table below presents our plan for the number of items to develop for the 2012-2013 pilots. Because the time frame for developing more than 10,000 items is very brief and because it is important to mitigate the risk associated with having educators (who may have other activities or priorities during the summer) write items, the Collaborative's proposal is based on the assumption that the Collaborative will develop 50 percent of the items and that educators will develop the other 50 percent. We have based our plan on the number of items of each type presented in RFP Table 1. We have applied a 15 percent overage for vendor-written items and a 25 percent overage for educator-written items to allow for attrition during internal and committee editorial reviews.

Table 6: 2012-2013 Pilots Item Development Plan

		Vendor-Written Items			Educator-Written Items		
Content Area	**Item Type**	**Items Needed**	**Overage**	**Items to Write**	**Items Needed**	**Overage**	**Items to Write**
ELA	Selected Response	500	1.15	567	500	1.25	616
	Technology-Enhanced Selected Response	700	1.15	805	700	1.25	875
	Constructed Response	600	1.15	686	600	1.25	749
	Technology-Enhanced Constructed Response	750	1.15	861	750	1.25	931
	Performance Task	53	1.15	56	52	1.25	70
	Technology-Enhanced	53	1.15	63	52	1.25	56

		Vendor-Written Items			Educator-Written Items		
Content Area	**Item Type**	**Items Needed**	**Overage**	**Items to Write**	**Items Needed**	**Overage**	**Items to Write**
	Performance Task						
	Stimuli			295			295
Math	Selected Response	500	1.15	567	500	1.25	616
	Technology-Enhanced Selected Response	850	1.15	973	850	1.25	1057
	Constructed Response	300	1.15	343	300	1.25	371
	Technology-Enhanced Constructed Response	850	1.15	973	850	1.25	1057
	Performance Task	53	1.15	56	52	1.25	70
	Technology-Enhanced Performance Task	53	1.15	63	52	1.25	56
Total Items		5262		6013	5258		6524

2. Conduct Cognitive Labs to Try Out New Types of Stimulus Materials, New Item Types, and Performance Tasks.

The cognitive labs and small-scale trials that comprise Part 3 of this project must inform the development of items for the pilot test. The timing of these activities is critical—the pilot test will occur in the spring of 2013 and the pilot-tested item and tasks will form the foundation of the SBAC assessment. Those items and tasks will form the anchors for the rest of the items to be developed for the 2014 field test, and therefore the integrity of the scale will depend on the pilot-tested items.

The significance of the research tasks expands as SBAC strives to improve the state of the-art of educational measurement. Integrating technology to achieve better measures of traits that have been difficult to measure in the past requires innovation, and necessarily carries risk of interim failures as ideas, beliefs, and conjectures are challenged and refined. Raising the bar higher, SBAC envisions that these objectives will be met with a financially sustainable model in which substantial parts of the test are automatically scored, either alone or in conjunction with some human process.

Put briefly, the cognitive lab studies and the small-scale trials carry the burden of refining and validating some of SBAC's most important innovations. These critical activities must be completed in time to inform the development of innovative items for the pilot test. Stretching timelines as far as reasonable, this allows at most about seven months from contract award until the last of the findings are available.

Success in this endeavor requires careful planning, vast resources, and a known mechanism for delivering interactive items in the lab and in the field (the small-scale trials will involve over 30,000 students). Our team offers a solution that will meet these challenges, while at the same time preserving open competition and advancing SBAC's intention to develop industrial-strength open-source solutions.

We begin with the recognition that some research questions are integral to the development of items, while others are more peripheral to the development process. For example, the general approaches and item types adopted in the item specification to get at harder-to-measure deeper knowledge form a nexus of item development. Overlooked implications or unexpected student reactions could lead to changes in the overall approach to item development. Similarly, new item types, interaction types, and other technology tools brought to bear may affect accessibility or cognitive load, have other unintended consequences, or fail to deliver the expected benefits.

The plans that are core to item development must be evaluated first. Cognitive lab studies can detect big general effects, and provide an opportunity to gather in-depth (if not representative) validity evidence about new approaches to measurement. Therefore, we plan to begin the cognitive lab studies almost immediately upon contract award. Findings from the cognitive labs will be available in time for the beginning of large-scale item development. Their iterative nature will provide an opportunity to detect the need for any substantial shifts in item specifications, and to test those adjustments in subsequent labs.

The small-scale trials, introduced immediately below, will be used to confirm key findings from the cognitive labs, and to test the efficacy of different presentation and stimulus-type options. While there is some risk of required re-work if the trials reveal that some cognitive lab findings do not survive the test of a more representative diverse trial, we have designed the cognitive labs to reduce this risk. The small-scale trials will also provide systematic tests of a variety of presentation options, allowing SBAC to make informed decisions about the composition (in terms of item types) and presentation of the tests. Many of these questions do not affect item development, but rather item delivery, for which there is a slightly longer time frame.

Delivering items both in the lab and in the small-scale trials requires a capable school-friendly test delivery system. To meet this requirement in a very short time frame, we propose to use AIR's successful test delivery system, which is currently in use in three SBAC states, and which requires little training and setup.

Our engine works by delivering low-bandwidth XML and media to browser-based item renderers, which are written in JavaScript. AIR is willing to open-source/open-license our item renderers, enabling any entity to delivery these items as they were delivered in the small-scale trials. The renderers handle the presentation, interactivity, and response capture for the items. This contributes to the SBAC goal of developing industrial-strength open-source solutions.

The open-sourcing of AIR item renderers will enable developers of the SBAC item authoring tool to integrate these renderers directly. Where the renderers diverge from SBAC preferences or specifications, developers of the authoring tool or SBAC test delivery system may modify them to meet SBAC specifications. These renderers are all HTML-5 compatible, and highly accessible.

This open-source strategy helps reduce timeline risk because these renders exist and are mostly currently in operational use (our simulation renderer will be used in operational statewide tests this spring). Use of these renders will ensure interoperability, and will encourage competition (because they will be open-source and unrestricted), and have already proven to minimize bandwidth requirements, to be easily integrated with multiple systems, and reduce the costs of developing technology-enhanced items. We integrate the same renders (basically, stand-alone Javascript modules with a well-defined API) with our test delivery system, our test development system, and our rangefinding system. SBAC contractors should have no problem with similar integration.

Using this approach, AIR can use our existing but proprietary test engine to deliver the small-scale tryouts (which will necessarily involve tens of thousands of students). This system is proven in statewide high-stakes operation. By open-sourcing the item renderers we eliminate any competitive advantage that this use of our system might bestow upon AIR. We believe that our renderers enable test developers to develop more cognitively challenging, problem-solving types of items. Enabling all testing organizations to develop and deliver these types of items supports our not-for-profit mission of using the best science to improve people's lives—in this case through the improvement of education and assessment.

3. Conduct small-scale trials to try out new stimulus materials, new item types, and automated scoring algorithms.

The objectives for the small-scale trials described in the RFP can be categorized under three general themes:

- To gather the student responses needed to inform scoring, both automated and human,
- To “try out” the tasks to get early indications of whether findings from the cognitive labs hold true with larger more diverse samples and whether students generate the range of responses expected, and
- To systematically evaluate presentation questions, including comparison of different layouts and formats for items and stimuli, and evaluating the impact of different mixes of item and stimulus types on tests.

Our plan meets the first goal, expanding the number of responses received on constructed responses for the automated scoring engines. The trials will provide the data for these studies, but we discuss the studies themselves under Part 2, the development and evaluation of automated scoring engines.

Our general approach to the small-scale trials is outlined under the previous objective. In general, we propose that the small-scale trials be designed as a series of embedded experiments to test specific hypotheses about how different item types, stimulus formats, layout presentations, and other factors work. The small-scale trials will provide a more systematic test of some of the findings from the cognitive labs, and investigate whether the items, and particularly the innovative items, function consistently across identifiable groups.

Recall that our strategy is to move forward with item development based on results from the cognitive labs. There is some risk that findings from the small-scale tryouts may identify some differential item function across groups or fail to uphold the cognitive lab findings.

Our plan accommodates potential findings contrary to those from the cognitive labs. In these cases items may require modification during the review process—an opportunity that we will preserve. Readers should also recall that development of items and stimuli that depend most on findings from the small- scale tryouts will be delayed until those results are available. This plan strikes the balance necessary to exploit the potential of these research opportunities and maintain SBAC’s pilot test schedule.

The final set of questions about the presentation of items and stimuli has at least four aspects:

- The particular media (format or type) used to present stimuli
- The layout of the presentation
- The effect of mixing item and stimulus formats on a single test
- The interaction of the stimulus format and available accessibility tools.

To address these questions we propose to develop different variants of items and stimuli, varying layouts and media in a true experimental design. This will allow us to detect factors that influence student performance, and therefore validity of the items.

Under Part 3 of this proposal we outline a draft plan for sampling over 30,000 students to participate in the small-scale trials. As part of this study, we embed a critical validity study. We will ask teachers to rate each student on each of the claims being tested by the various items. Along with other evidence, this evidence will be used to validate the belief that student responses on particular items provide evidence of particular claims outlined in the content specifications. Details appear under Part 3 of this proposal.

4. Design controlled studies of three item procurement options (state managed, state submitted, and SBAC managed1) to determine the relative merits of each option.

Leveraging various procurement methods and existing assessment content can be an important factor in the cost-effective development of the SBAC balanced assessment system. CTB has partnered with HumRRO, an organization with extensive experience in conducting independent evaluation studies, to carry out this study to provide SBAC with an evaluation of the various procurement options. HumRRO currently serves as independent evaluator for several important efforts, including the National Assessment of Educational Progress (NAEP). HumRRO routinely conducts special studies for the NAEP program to investigate areas of particular challenge, including a recent effort that examined the quality of NAEP mathematics items. The role that HumRRO plays on the NAEP program and the expertise staff has developed from conducting evaluations of other education programs are directly relevant to the proposed SBAC procurement option evaluation study. While CTB will play a role in implementing the study and collecting item samples and review data, HumRRO will independently carry out all data review and evaluation procedures.

The evaluation study will begin with the definition of implementation parameters. For the purposes of this proposal, we have assumed that samples of items will be obtained from each of the three procurement channels. Items developed under the SBAC-managed option will be evaluated from the total item development described under Task 1 of this proposal. Items from the state-managed option will be procured from state-managed development efforts specifically undertaken for SBAC in two states recruited by SBAC, in accordance with the Q&A responses to Q66 and Q68. For the state-ubmitted option, items will come from other sources within states preliminarily identified by SBAC.

The first activity of the study will be to define the definitions and implementation parameters for obtaining item samples. CTB and HumRRO will consult with SBAC to determine the best configuration of the study and plan for the acquisition of the three samples to ensure fair comparisons while maintaining congruence with the SBAC guidelines for state participation in item development.

HumRRO recommends the use of random assignment when conducting the small-scale study to compare procurement options with sampling occurring at the state level. Our recommendations for the detailed design of the sampling plan are described under Task 4. We will discuss with SBAC leadership the possibility of recruiting more than two states to participate in the study. For two states, we recommend collecting 1200-1500 items from each state participating in any of the three options. From this overall pool, HumRRO will select a smaller sample for analysis.

We are recommending an aggressive implementation timeline to allow for the study items to be part of the 2012–2013 Pilot Test. This will allow pilot test data, along with results from independent item reviews, to be used in the evaluation study. The evaluation methodology will include procedures to document the item development processes implemented under each option, and will identify outcome measures of cost and quality. We will collect as much data on implementation as possible by attending or observing pre-planned development meetings.

For outcome measures, HumRRO will independently examine the time and costs required to develop items under each option, including the time and costs associated with item revision and refinement, and the resulting quality of the items that are developed. To assess item quality, prior to the Pilot Test administration, the items will be rated as part of the Task 1 item reviews on a number of item quality parameters, including standard alignment, cognitive rigor, and technical quality. We propose to randomly select items from each of the three procurement options and provide them to the experts to rate; we anticipate this will occur during one of the scheduled content reviews. Proportions of item types (multiple choice, constructed response, technology enhanced) will mimic targets for the operational assessment. Once field test data are available, we will examine classical item statistics and flag items whose functioning appears questionable.

The final evaluation report, produced independently by HumRRO, will describe evaluation procedures, outcomes, and recommendations that highlight strengths to continue as well as improvements to enhance the process.

5. Develop and test automated scoring models/algorithms for appropriate items/tasks.

Under this activity SBAC seeks to:

- Understand and extend the existing state of the art in automated scoring.
- Identify, extend, and evaluate various open-source scoring engine.
- Based on the abilities of automated scoring approaches demonstrated under this contract, develop a vision for scoring student responses from the pilot-test, field-test, and operational SBAC assessments. The resulting vision will likely include specification of automated scoring approaches, validation mechanisms, and integration with human scoring where necessary.

These are ambitious objectives, and SBAC hopes to achieve them on a very aggressive timeline. The CTB/AIR team offers a plan that will meet these timelines and objectives. The plan will

- support immediate work on the adaptation or development of open-source automated scoring solutions
- make proprietary automated scoring solutions available for evaluation
- allow for refinement and validation of those solutions based on data collected during the small-scale trials
- optionally, help SBAC establish an organization and community of developers to support future development and sustainability of the open-source solutions.

The response to questions about the RFP stated that the deliverable is specifications for the scoring engines. In this proposal, we propose to delivery open-source engines, which will include documentation (including specifications). This should help SBAC advance the objective of developing open-source assessment solutions.

We propose to begin work on the scoring engines immediately. The timeline does not allow this work to await the availability of data from the small scale trial. While we will use that data when it becomes available, we plan to get a jump-start on this activity using data gathered from SBAC states. We propose to begin the work with existing student responses solicited from SBAC states. We recognize that many responses will have to be transcribed from paper, with the transcriptions preserving salient errors.

Working with SBAC, AIR will identify four open-source solutions to serve as a basis. Each solution will either be an existing scoring engine, or (for example) natural language processing tools possessing the features and capabilities needed for a scoring engine.

Under Part 2 of this proposal we describe the current state-of-the-art and available open source tools. There, we describe how we will go about developing and testing various solutions, including the delivery of up to four open-source scoring engines.

Part 2 also describes the iterative validation and revision process that will be used to evaluate the performance of each iterative version of the engines. Much of this work will take place using existing student responses before the small-scale pilot. The process will continue after the small-scale pilot data become available.

In addition to the open-source solutions, AIR offers some of our proprietary solutions for evaluation. In each case we are willing to negotiate terms under which we can make these available under an open-source license, should SBAC choose to go forward with our proprietary solutions.

Table 7 summarizes the proposed scoring engines and our strategy to ensure continuing competition where open-source solutions are not feasible. Some of the details (such as particular open-source tools mentioned) are discussed at more length in Part 2 of the proposal.

Table 7: Summary of Scoring Engines to be Developed and Evaluated

Item type or response type	Open Source solution	Proprietary Solution (optional)	Strategy to ensure continued competition
Selected response, including drag and drop, hotspot, etc., as supported by QTI	Scoring engine based on JQTI	N/A	N/A
Essay or other long text	Up to 2 black-box engines based on S-Space or SemanticVectors	Autoscore, others SBAC may negotiate to evaluate	AIR is willing to negotiate open-source licensing for our proprietary engines, or offer commercial licenses to any other entity
Other natural language constructed response		Proposition Scorer, others SBAC may negotiate to evaluate	
Graphic response	N/A	Graphic Response Item Scorer	
Equation response	Equivalence-matching engine based on open-source symbolic processor	N/A	N/A

6. Produce reports in SBAC determined formats regarding:

a. Improvement and refinement of each procurement option.

b. Results of cognitive labs.

c. Results of small-scale trials including automated scoring trials.

d. As necessary for high quality item development, recommendations for changes/enhancements or appendices to SBAC content specifications, item/task specifications, stimulus specifications, bias and sensitivity guidelines, and accessibility guidelines.

e. As necessary, to maintain consistency and quality of item/task production proposed, recommendations for changes/enhancements or appendices to

i. Training materials for item/task/stimulus writing/development including recommendations for improvement of item/task/stimulus writer/developer training materials.

ii. Training materials for content reviews, bias and sensitivity reviews, and accessibility reviews including recommendations for improvement to training materials.

iii. Procedures and materials for content reviews, bias and sensitivity reviews, and accessibility reviews including recommendations for improvement to reviews.]

As discussed in section L1-19, we will gather both formal and informal feedback from users of the training materials, item specifications, and review guidelines and protocols throughout the item/task/stimulus development cycle. In addition to gathering feedback from participant users of the materials, feedback obtained from the results of the cognitive labs and small-scale trials will be compiled and submitted to SBAC for consideration and for use in making appropriate changes or enhancements to the training materials and/or item/task/stimuli specifications.

We recommend that participants' feedback on training materials or specifications be gathered at three points in the development or review process. We recommend this time sampling approach because we anticipate that participants' feedback may evolve over time as they gain experience with the use of item/task/stimulus specifications or with review guidelines and processes. As writers and reviewers delve deeper into the specifications and guidelines and internalize the information, their understanding of the information will evolve, and we would expect that the nature of their feedback may also evolve.

Item/Task Development Feedback

- After item/task writing training, from both item writers and editors (to focus on clarity of item writing training and trainers, etc.)
- After initial round of item batches, from both writers and editors (to focus on specifications and other writing materials, item review and revision process, effectiveness of editor feedback, etc.)
- At the conclusion of the item writing phase (to focus on overall item development process, materials, schedule, etc.)

Content, Bias/Sensitivity, Accessibility Review Feedback

- After item review training (to focus on the clarity and effectiveness of training materials and procedures, etc.)
- At a midpoint during the review period after a few workflows of items have been reviewed
- At the conclusion of the item reviews, to focus on overall review processes, guidelines, and effectiveness of training materials and modules, workflow procedures, etc.

In addition to gathering feedback from participants about materials and processes, the content lead and editorial staff of the Collaborative partners will provide feedback. We will gather feedback in both formal and informal ways. Formal methods may include surveys and focus groups. Informal methods may include annotations and comments that they mark on PDF copies of the item specifications, review guidelines and protocols, and training materials as they use them when working with item writers or reviewing and editing items. As noted elsewhere, the College Board will review a sample of grade 8 and high school items to determine whether the items will provide evidence toward determining students' college and career readiness. The College Board's review will provide another source of feedback on the item/task/stimulus specifications and item writer and review training materials.

7. Develop item/task/stimulus per SBAC participation policies. Hire and train teachers/educators from SBAC states as item/task/stimulus writers/developers using SBAC-developed item/task/stimulus writer/developer training materials.

The Collaborative understands the goals for this project, and we are qualified and prepared to partner with SBAC to achieve those goals. Collectively our team's work as primary full-service vendors for statewide educational assessment projects, together with our success in providing item development and review services for our clients, positions the Collaborative well to ensure accurate high-quality vendor- and teacher-developed items.

Our extensive experience in assessments involving teacher-developed items encompasses the development and review of the all item types to be developed for the SBAC. For example, in Michigan where items are developed by teachers, DRC's role currently includes training teachers to develop items online using state-approved item and test specifications. DRC's role also includes providing feedback online, as needed. For Nebraska, teachers also develop items for the statewide assessment using item-writing templates. DRC's role in the teacher-item development process involves the review of the items with feedback provided to the Nebraska Department of Education. For the Commonwealth of Pennsylvania, teachers are involved in the development and review of performance tasks included in units and lesson plans. DRC's role includes monitoring teacher feedback provided online. Based upon teacher feedback, performance tasks can be revised and reposted electronically for use in classrooms throughout the Commonwealth.

Similarly, CTB has conducted multiple trainings in the art and practice of item writing for many of its former and current customers, including Missouri, Maryland, North Dakota, Qatar, and Bermuda. Each of these customers has used and/or continues to use educator-developed items and tasks in their state- and district-wide assessments. Teachers' feedback on surveys at the conclusion of item writing training workshops is always overwhelmingly positive in terms of the professional development experience and the new skills the educators acquired. CTB has also developed an on-line tool in its proprietary Acuity system. Using this tool, educators can create their own performance tasks; our customers in New York City have received training and have begun successfully using the tool.

Our Collaborative partners, all with proven success nationally, affirm that the training of vendor and educator writers for the SBAC's items, tasks, and stimuli will adhere to the processes and policies established by the consortium's governing body, and that the training materials to be used will have been developed according to the consortium’s specifications. Our combined expertise in working with teachers to develop items will provide a solid foundation for the effective implementation of SBAC item/task specifications and training materials.

Our team has decades of experience providing stimulus and item writing training and item reviews that follow standardized procedures and use detailed training materials; furthermore, each partner has successfully demonstrated the ability to customize training and materials for its customers’ needs, including projects requiring alignment with the CCSS.

Hire and Train Stimulus Writers/Developers

We understand that the SBAC 08 RFP deliverables include the development of participation policies and procedures for SBAC member states and stakeholders to be involved in SBAC assessment development activities as well as recruitment and selection criteria for item/task/stimulus writers/developers. Based on the response to bidders' questions Q9, we also understand that a deliverable of SBAC 08 is the identification of qualified individuals to serve as stimulus writers. As mentioned previously, given the very short time frame for authoring or searching for and selecting high-quality ELA stimuli that meet the stimulus specifications, our proposal is based on the assumption that the Collaborative will select or author the stimuli for the cognitive lab and small-scale trials and the majority of the stimuli for the 2012-2013 pilots. We are ready to start immediately with selecting and authoring stimuli using our experienced writers. We will supplement our cadre of stimulus writers with writers identified by the SBAC 08 contractor.

The Collaborative will train all qualified stimulus writers to perform assigned stimulus development tasks. Training will be provided to ensure that these individuals are selecting or developing stimuli that reflect the spirit and requirements of the common core standards. Based on the draft list of specifications topics for ELA stimuli, we anticipate that the training materials developed by the SBAC 08 vendor will address general requirements of the specifications as well as more in-depth training in requirements by grade level and text type. We anticipate that training will also address requirements for text complexity, text features, and text structure by grade level and text type. Our assumption is that all stimulus writers will participate in training that addresses general requirements as well as requirements that are common across grade levels, including instruction on how to use the online authoring system to submit stimuli and document metadata, including source documentation. We assume that in-depth training in developing stimulus materials that will support CCSS-aligned items will be organized by grade span. We recommend that as part of the training, writers be instructed to also create and submit text maps which will outline the theme or main ideas, craft and structure elements, and other key elements of the text that will support CCSS-aligned items.

For the purpose of estimating costs associated with training stimulus writers, we assumed that each writer would participate in not more than half a day of training (3.25 hours). Our price proposal is based on the assumption that all training is conducted via online webinars. While we have assumed that the stimulus writers will participate in live webinars, the webinars will be recorded and be available online for review at the writer's convenience (without additional payment for training time).

Also, for costing purposes, we assume that the time to develop one ELA stimulus ranges from half a day to 1.5 days, depending on grade level, stimulus type, and stimulus length. For equity purposes and for estimating the cost of creating commissioned stimuli, we propose to pay stimulus development participants based on a flat daily rate per reading passage based on text length. Our price proposal is based on an average of one day per stimulus. We have also assumed that stimuli for language items, such as brief paragraphs for assessing editing skills, will be developed as part of the item and included in the payment for the development of the item.

Following the training session, we will obtain participants' feedback on the quality of the training, using the feedback and quality control guidelines and procedures developed by the SBAC 08 vendor. We will also obtain stimulus writers' feedback on the stimulus specifications after they have had experience using the specifications. We will keep track of questions or comments that participants submit as they use the specifications and aggregate and submit the feedback to SBAC. Based on the Collaborative's use of the stimulus specifications to create or select stimuli for the cognitive lab and small-scale trials items, we will provide feedback to SBAC for consideration. As part of the first management meeting soon after contract award, we will discuss the means and mechanisms for using SBAC procedures for providing feedback on specifications and guidelines.

Stimulus writers will be given specific writing assignments based on the needs of the blueprints and item development plan. The assignments will be sent to the writers via electronic workflows within the interim item authoring system. The writer will also be given a due date for the completion of the assignment. The SBAC Stimulus Specifications document will be accessible as a PDF within the item authoring system. The Collaborative's assessment editors will be available via telephone and email to respond to writers' questions and provide support in the use of the online system. Our proposed plan for the development or selection of ELA stimuli is discussed in detail in section L1-12.

Hire and Train Item/Task Writers/Developers

Based on the response to bidders' questions Q9, we also understand that a deliverable of SBAC 08 is the identification of qualified individuals to serve as item/task writers. We anticipate starting item/task development for the 2012–2013 pilots in June 2012. Our assumption is that a sufficient number of qualified individuals, based on our estimates of the number of writers needed, will be recruited and screened by the SBAC 08 vendor by early May 2012.

Our assumption is that selection criteria include previous experience as assessment item writers or a process for identifying individuals with the aptitude to become strong writers. CTB will manage the process of contracting with the qualified participants beginning in early May and all writers should be on board before the end of May. In section L1-12 below, we provide details of our item/task development plan, including our estimates of the number of item/task writers needed based on our proposed schedule.

The Collaborative will use the training materials developed by the SBAC 08 vendor to train the item/task writers. We anticipate that the training materials developed by the SBAC 08 vendor will address general requirements of the specifications as well as more in-depth training in requirements by grade level and text type. We assume that the item writer training will focus on the SBAC item specifications and requirements for each item format and not primarily on the basics of writing quality test items.

We anticipate that the training will be composed of modules, with some modules focusing on general requirements, such as how to use the item specifications and style guide and how to use the interim item authoring system (which the CTB will develop). Other modules will be more sharply focused on item formats and on item specifications for a grade level or grade span. We also understand that a major aspect of training will be the accessibility and accommodation guidelines and the addressing of accessibility from a construct perspective, so that writers are trained to create accessible items from the outset, rather than needing to retrofit items later on. Our price proposal is based on the assumption that each item/task writer will participate in one day (6.5 hours) of training; the training modules may occur over several days rather than on a single day. Further, our price proposal is based on the assumption that all training will be conducted via live online webinars. The webinar training modules will be recorded and available online for review by the writers as needed (without additional payment for training time).

The Collaborative has extensive experience in training both vendors and educators in how to develop high-quality items for a project. We have included sample materials in Appendix C to illustrate our typical methodology and how our experience aligns well with the item development requirements of this RFP. Appendix C includes a portion of a PowerPoint slide deck template CTB uses when training our item development vendors. Prior to its use for any project, it is customized with appropriate item examples, art guidelines, key item development points, sample rationales, and the like. The level of detail of this training is determined by the level of experience with the vendor and the writers' expertise. We also provide recordings of all the live training sessions for the item writers to reference at their convenience. Our assessment editors encourage item writers to contact them at any time during the project’s development cycle in order to achieve continuous improvement. The Collaborative will bring our experience and expertise to this project and the use of the SBAC training materials and processes, thus ensuring that participants are well trained and have a positive, productive experience.

Item/task writers will be given specific item writing assignments based on the needs of the blueprints and item development plan. The assignments will be sent to the writers via electronic workflows within the interim item authoring system. A single assignment may range from 10 to 25 items. The writers will also be given a due date for the completion of each assignment. The SBAC Stimulus Specifications document will be accessible as a PDF within the item authoring system. The Collaborative's assessment editors will be available via telephone and email to respond to writers' questions and provide support in the use of the online system.

In order to estimate the item/task development budget for items/tasks written by SBAC participants, we estimated the amount of time to draft each type of item and make revisions based on an editor's feedback. The amount of time per item is based on our experience with novice item writers and our estimates based on our examination of the sample SBAC item specifications. For equity purposes and in order to make reasonable estimates of the budget for item/task development, our price proposal is based on the assumption that writers will be compensated based on the estimated number of items per day, as shown in the table below, regardless of how much time they actually spend. Based on the responses to bidders’ questions, we used an approach based on a per diem rate; however, we can discuss alternative compensation models with SBAC.

Table 8: Estimated Item Production Rate for Educator-Written Items/Tasks

Item Type	Estimated Hours per Item	Items per Day
Selected Response	1	6
Selected Response - Technology Enhanced/Enabled	2	3
Constructed Response	3	2
Constructed Response - Technology Enhanced/Enabled	4	1.5
Performance Task	6.5	1
Performance Task - Technology Enhanced/Enabled	6.5	1

In order to estimate the number of participants needed to develop half the items for the 2012-2013 pilots (6,524), we assumed that each writer would be responsible for approximately 50 items and 2 tasks. In section L1-12 we describe how we propose to organize writers into teams of three, with a total of 21 teams per content area. Hence, our proposal is based on the assumption that the Collaborative will contract with 126 participants to write items/tasks.

8. Hire and train teachers/educators to serve as content reviewers using SBAC-developed content review training materials. Selection of content reviewers will be based on SBAC participation policies and SBAC writer/developer qualifications.

One of the overarching goals of the SBAC balanced assessment system is to measure student achievement against the CCSS, which provide expectations for college- and career-readiness (CCR). To further that goal, the Collaborative will leverage the expertise of the College Board during the item development process. Based on its extensive experience in helping students prepare for college and determine college readiness, the College Board will review a sample of items at the secondary level to determine whether they will provide sufficient evidence of progress toward SBAC CRR goals. The College Board has extensive experience in this area and will bring its experience to the review of these items.

The following activities will be part of this review:

- Guidelines and materials for CCR review (checklists, guidelines, etc.) will be developed and made available to item writers at grade 8 and HS to inform item development.
- During item development, as items/tasks are finished, a sample of items/tasks at grade 8 and HS for both ELA and Mathematics will be randomly selected for review by the College Board. The samples will include approximately the same distributions of selected-response, constructed-response, and performance task (and technology-enhanced items) as the full development.
- The College Board will review the sample of items during item development and before the Content, Bias/Sensitivity, and Accessibility reviews. The College Board will document comments on individual items, along with other reviewer comments in the authoring system. CTB will provide summaries of these comments back to the grade 8 and HS authoring teams.
- The College Board will also provide a brief final report of overall comments and recommendations based on this audit. These recommendations can provide input to the content review committees at the appropriate grade levels.

We understand that the SBAC 08 RFP deliverables include the development of participation policies and procedures for SBAC member states and stakeholders to be involved in SBAC assessment development activities, and that the deliverables include criteria for content reviewers. Furthermore, we understand that recruiting participants for the committees is within the scope of work of the SBAC 08 contract. Therefore, our assumption is that SBAC and the 08 vendor will recruit, screen, and select qualified participants for the stimulus and item content review committees, based on the number of participants we identify as being needed. We assume that the content reviewer recruitment and selection process will occur in spring 2012 so that stimulus reviewers are available immediately and that item reviewers are selected by the time they are needed in early fall 2012. CTB will manage the process of contracting with the reviewers to perform the review tasks.

The CTB program manager will use the change management process to adjust the scope and price or also the schedule as needed if the number of content review participants needed, how participants review items, or the amount of the stipend paid to participants differs from the assumptions we have made Our commitment is to perform the requirements of the scope of work in a manner that will yield quality assessment items.

We will use the stimulus and item content review committee and facilitator training materials developed by the SBAC 08 vendor to train reviewers. We will also train stimulus and item content reviewers in the use of established protocols for recording their responses to the content review of items and capturing their recommendations for item revisions.

Our price proposal is based on the assumption that stimulus or item content review committee participants will attend an online live webinar training session of approximately two hours' duration. Training for stimulus review and item accessibility review will be in different modules and will be conducted separately.

Following the training, we will implement the quality control guidelines and procedures developed by the SBAC 08 vendor to ensure that reviewers are adequately trained and able to perform their duties. We will also implement the criteria and procedures for record-keeping, which are also developed by the SBAC 08 vendor. Our technical and price proposals are based on the assumption that all stimulus and item reviews will occur as asynchronous online reviews. That is, reviewers will be sent an electronic workflow within the interim item authoring system containing a queue of stimuli or items to review, and will be given a due date for completing their individual review of the stimuli or items. Each reviewer will work independently (as opposed to coming together in a live online webinar to discuss stimuli or items), and submit their ratings and comments via the interim online authoring system. The Collaborative's assessment editors will be available via telephone and email to respond to reviewers' questions and provide support in the use of the online system and the SBAC review guidelines. The interim item authoring tool that we propose to use will capture, track, and aggregate reviewers' ratings and comments. CTB will compile reports summarizing the review data, and will submit reports to SBAC in preparation for the joint review and reconciliation sessions.

Stimulus Content Reviewers

Our proposal is based on the assumption that three reviewers per grade level will be needed, for a total of 21 ELA stimulus reviewers. We have assumed that three reviewers will each review all stimuli for the 2012-2013 pilot items for the grade level. Based on the responses to bidders' questions, we understand that SBAC anticipates approximately 800 total ELA stimuli, of which we estimate that 590 will be for the 2012-2013 pilot items and will be reviewed by committees of educators. Please refer to our discussion of the number of stimuli needed in section L1-12.

Based on our experience, we estimate that five stimuli can be reviewed per hour, which includes the time to read the stimulus and enter specific comments or suggestions into the interim online system. For equity, each reviewer will be paid a flat daily rate based on 32 stimuli per day, regardless of how many hours a reviewer actually spends to complete the review of the assigned stimuli.

To summarize, our assumptions regarding the hiring and training of ELA stimuli reviewers are:

- Three reviewers per grade; 21 reviewers total
- Two hours of live online webinar training per reviewer
- Each reviewer reviews all stimuli for a grade (approximately 84) over a period of a few weeks
- Each stimulus is reviewed by three reviewers
- Average review rate of five stimuli per hour
- Reviewers are paid a flat daily rate based on 32 stimuli per day based on a 6.5 hour day
- Reviewers work individually and will most likely log into the system from home during evening or weekend hours.

Item Content Reviewers

For the 2012–2013 items, we determined that 84 ELA and 84 mathematics item content reviewers will be needed, for a total of 168. The item reviewers will be organized based on content area and grade level with two teams of six reviewers per grade/content. We propose that six content reviewers review each item and that each reviewer will be routed batches of items to review over a period of six to eight weeks. Based on our experience with face-to-face content review meetings where committees discuss each item individually and propose specific edits, we know that, on average, a committee can review 10 items per hour. We estimate that reviewers can review 20 multiple choice or constructed- response items per hour or two performance tasks per hour because participants will be reviewing items individually online. Reviews should proceed more quickly than in traditional committee meetings. Therefore, our price proposal is based on the assumption that each reviewer will be contracted for five days based on a 6.5 hour day. For equity, based on these assumptions, we will pay a flat rate based on 130 selected- or constructed-response items per day or 13 performance tasks per day, regardless of how many clock hours a reviewer actually spends completing the assignments.

We determined the number of content reviewers needed and the cost in our price proposal by making the assumptions listed below.

- Content reviews occur in September and October 2012
- Reviewers will work individually to review items and will most likely log into the system from home during evening or weekend hours
- Each reviewer will review items intermittently over several weeks for a total of 32.5 hours (five days)
- Each reviewer will be contracted for a flat rate based on the daily item review rate plus two hours of training, based on a 6.5 hour day
- Each multiple-choice or constructed-response item, on average, will take three minutes to review and to document the review response
- Each performance task, on average, will take 30 minutes to review and to document the review response
- Six reviewers will review each item, working individually.

Table 9: Item Content Review Participants Needed

		Number of items for 2012-2013	**Total items**	**Items to Review if each is reviewed by 6 people**	**Review Rate per Person per Day**	**Total Number of Review Days**	**Number of Participants if 12 per grade**	**Review days per Participant**
ELA	Selected Response	1,183	6,090	36,540	130	281	84	5
	Technology-Enhanced Selected Response	1,680						
	Constructed Response	1,435						
	Technology-Enhanced Constructed Response	1,792						
	Performance Task	126	245	1,470	13	113		
	Technology-Enhanced Performance Task	119						

		Number of items for 2012-2013	Total items	Items to Review if each is reviewed by 6 people	Review Rate per Person per Day	Total Number of Review Days	Number of Participants if 12 per grade	Review days per Participant
Math	Selected Response	1,183	5,957	35,742	130	275	84	5
	Technology-Enhanced Selected Response	2,030						
	Constructed Response	714						
	Technology-Enhanced Constructed Response	2,030						
	Performance Task	126	245	1,470	13	113		
	Technology-Enhanced Performance Task	119						
TOTALS		12,537	12,537	75,222		782	168	

9. Hire and train qualified individuals to serve as bias/sensitivity reviewers using SBAC- developed bias/sensitivity review training materials.

The members of the Collaborative bring extensive experience in conducting bias/sensitivity reviews for stimulus materials and assessment items, with tens of thousands of items being reviewed annually and hundreds of educators trained to perform reviews. During these formal reviews, all individual test items will be evaluated using the SBAC review guidelines to identify items that may be biased or material that may be of a sensitive nature, and for overall fairness. The test items, tasks, and stimuli must avoid potential sources of bias related to gender, ethnic, age, or other stereotypes, and avoid language, symbols, words, phrases, or examples that reflect a gender or ethnic bias or that are otherwise potentially offensive or inappropriate for any sample of the student population.

We understand that the SBAC 08 RFP deliverables include the development of criteria for recruiting and selecting bias/sensitivity reviewers. Furthermore, we understand that recruiting participants for the committees is within the scope of work of the SBAC 08 contract. Therefore, our assumption is that SBAC and the 08 vendor will recruit, screen, and select qualified participants for the bias/sensitivity review committees based on the number of participants we identify as being needed. We anticipate that the bias/sensitivity reviewer recruitment and selection process will occur in spring 2012 just as soon as the participation policies and procedures are ready for use. Further, we assume that stimulus and item reviewers are selected and on board by the time reviews need to begin according to the proposed schedule. The Collaborative will contract with the reviewers to perform the review tasks.

The CTB program manager will use the change management process to adjust the scope and price or also the schedule as needed if the number of bias/sensitivity review participants needed, how participants review items, or the amount of the stipend paid to participants differs from the assumptions we have made Our commitment is to perform the requirements of the scope of work in a manner that will yield quality assessment items.

We will also use the bias/sensitivity review committee and facilitator training materials developed by the vendor for the SBAC 08 contract to train reviewers in how to use the bias/sensitivity review guidelines and criteria to review ELA stimuli or test items for potential sources of bias or sensitivity. We will also train bias/sensitivity reviewers in the use of established protocols for recording their responses to the bias/sensitivity review of items and capturing their recommendations for item revisions.

Our price proposal is based on the assumption that stimulus or item bias/sensitivity review committee participants will attend an online live webinar training session of approximately two hours' duration. Training for stimulus review and item accessibility review will be in different modules and will be conducted separately.

Following the training, we will implement the quality control guidelines and procedures developed by the SBAC 08 vendor to assure that reviewers are adequately trained and able to perform their duties. We will also implement the criteria and procedures for record keeping, which are also developed by the SBAC 08 vendor. Our technical and price proposals are based on the assumption that all stimulus and item reviews will occur as asynchronous, online reviews. That is, reviewers will be routed a workflow containing a queue of stimuli or items to review and will be given a due date for completing their individual review of the stimuli or items. Each reviewer will work independently (as opposed to coming together in a live, online webinar to discuss stimuli or items), and submit their ratings and comments via the interim online authoring system. The Collaborative's assessment editors will be available via telephone and email to respond to reviewers' questions and provide support in the use of the online system and the SBAC review guidelines. The interim item authoring tool that we propose to use will capture, track, and aggregate reviewers' ratings and comments. CTB will compile reports summarizing the review data and submit reports to SBAC in preparation for the joint review and reconciliation sessions.

Stimulus Bias/Sensitivity Reviewers

Our proposal is based on the assumption that four bias/sensitivity reviewers per grade level will be needed, for a total of 28 ELA stimulus bias/sensitivity reviewers. We have assumed that the four reviewers will each review all stimuli for the grade level. Based on the assumption of 590 ELA stimuli for the 2012-2013 pilot items, there will be an average of 84 stimuli per grade level. Based on our experience, we estimate that five stimuli can be reviewed per hour, which includes the time to read the stimulus and enter specific comments or suggestions into the interim online system. For equity, each reviewer will be paid a flat daily rate based on 32 stimuli per day, regardless of how many hours a reviewer actually spends to complete the review of the assigned stimuli.

To summarize, our assumptions regarding the hiring and training of ELA stimuli reviewers are:

- Four reviewers per grade level; 28 reviewers total
- Two hours of live, online webinar training per reviewer
- Each reviewer reviews all stimuli for a grade level (approximately 84) over a period of a few weeks
- Each stimulus is reviewed by four reviewers
- Average review rate of five stimuli per hour
- Each reviewer will be contracted for a flat rate based on the daily item review rate plus two hours of training, based on a 6.5 hour day
- Reviewers work individually and will most likely log into the system from home during evening or weekend hours

Item Bias/Sensitivity Reviewers

Our proposal is based on the assumption that 112 bias/sensitivity reviewers will be needed for the review of items for the 2012-2013 pilots. We assumed that reviewers will be organized by grade and content area, with eight reviewers per grade/content. We propose that four bias/sensitivity reviewers review each item and that each reviewer will review approximately half the items for the grade level. We estimate that reviewers can review 25 multiple-choice or constructed-response items per hour or two performance tasks per hour. Therefore, our price proposal is based on the assumption that each reviewer will be contracted for 4 days based on a 6.5 hour day. For equity, based on these assumptions, we will pay a flat daily rate based on 163 items or 13 performance tasks per day, regardless of how many clock hours a reviewer actually spends reviewing the assigned items.

We determined the number of item bias/sensitivity reviewers needed and the cost in our price proposal by making the assumptions listed below.

- Bias/Sensitivity reviews occur in September and October 2012
- Reviewers will work individually and will most likely log into the system from home during evening or weekend hours
- Each reviewer will review items intermittently over several weeks for a total of 4 days (26 hours)
- Each reviewer will be contracted for a flat rate based on the daily item review rate plus two hours of training, based on a 6.5 hour day
- Items can be reviewed at the rate of 25 items per hour, including documenting the review response
- Each performance task, on average, will take 30 minutes to review and to document the review response
- Four reviewers will review each item, working individually, and the reviewers for any given item represent a diversity of demographic constituencies

Table 10: Item Bias/Sensitivity Review Participants Needed

		Number of items for 2012-2013	Total items	Items to Review if each item is reviewed by 4 people	Review Rate per Person per Day	Total Number of Review Days	Number of Participants if 8 per grade	Review days per Participant
ELA	Selected Response	1,183	6,090	24,360	163	149	56	4
	Technology-Enhanced Selected Response	1,680						
	Constructed Response	1,435						
	Technology-Enhanced Constructed Response	1,792						
	Performance Task	126	245	980	13	75		
	Technology-Enhanced Performance Task	119						

		Number of items for 2012-2013	Total items	Items to Review if each item is reviewed by 4 people	Review Rate per Person per Day	Total Number of Review Days	Number of Participants if 8 per grade	Review days per Participant
Math	Selected Response	1,183	5,957	23,828	163	146	56	4
	Technology-Enhanced Selected Response	2,030						
	Constructed Response	714						
	Technology-Enhanced Constructed Response	2,030						
	Performance Task	126	245	980	13	75		
	Technology-Enhanced Performance Task	119						
TOTALS		12,537	12,537	50,148		445	112	

10. Hire and train educators and other qualified individuals to serve as accessibility reviewers using SBAC-developed accessibility review training materials.

Likewise, we understand that the SBAC 08 RFP deliverables include the development of criteria for recruiting and selecting accessibility reviewers and that recruiting participants for the committees is within the scope of work of the SBAC 08 contract. Therefore, our assumption is that SBAC and the 08 vendor will recruit, screen, and select qualified participants for the accessibility review committees based on the number of participants we identify as being needed. We anticipate that the accessibility reviewer recruitment and selection process will occur in spring 2012 just as soon as the participation policies and procedures are ready for use so that stimulus and item reviewers are selected and on board by the time reviews need to begin according to the proposed schedule. The Collaborative will contract with the reviewers to perform the review tasks.

The CTB program manager will use the change management process to adjust the scope and price or also the schedule as needed if the number of accessibility review participants needed, how participants review items, or the amount of the stipend paid to participants differs from the assumptions we have made Our commitment is to perform the requirements of the scope of work in a manner that will yield quality assessment items.

We will also use the accessibility review committee and facilitator training materials developed by the vendor for the SBAC 08 contract to train reviewers in how to use the accessibility review guidelines and criteria to review ELA stimuli or test items for accessibility by all students and especially those with disabilities. We will also train accessibility reviewers in the use of established protocols for recording their responses to the accessibility review of items and capturing their recommendations for item revisions.

Our proposal is based on the assumption that stimulus or item accessibility review committee participants will attend an online, live webinar training session of approximately two hours' duration. Training for stimulus review and item accessibility review will be in different modules and will be conducted separately.

Following the training, we will implement the quality control guidelines and procedures developed by the SBAC 08 vendor to ensure that reviewers are adequately trained and able to perform their duties. We will also implement the criteria and procedures for record-keeping, which are also developed by the SBAC 08 vendor. Our technical and price proposals are based on the assumption that all stimulus and item reviews will occur as asynchronous, online reviews. That is, reviewers will be routed a workflow containing a queue of stimuli or items to review and will be given a due date for completing their individual review of the stimuli or items. Each reviewer will work independently (as opposed to coming together in a live, online webinar to discuss stimuli or items), and submit their ratings and comments via the interim online authoring system. The Collaborative's assessment editors will be available via telephone and email to respond to reviewers' questions and provide support in the use of the online system and the SBAC review guidelines. The interim item authoring tool that we propose to use will capture, track, and aggregate reviewers' ratings and comments. CTB will compile reports summarizing the review data and submit reports to SBAC in preparation for the joint review and reconciliation sessions.

Stimulus Accessibility Reviewers

Our proposal is based on the assumption that two accessibility reviewers per grade level will be needed, for a total of 14 ELA stimulus accessibility reviewers. We have assumed and that two reviewers will divide the work of reviewing stimuli for a grade level, with each reviewer reviewing half the stimuli. Based on the assumption of 590 ELA stimuli for the 2012-2013 pilot items, there will be an average of 84 stimuli per grade level or 42 per reviewer. A reviewer may request that a second reviewer also review a stimulus; we made the assumption that 10 percent of stimuli will receive a second review.

Based on our experience, we estimate that five stimuli can be reviewed per hour, which includes the time to read the stimulus and enter specific comments or suggestions into the interim online system. For equity, each reviewer will be paid a flat daily rate based on 32 stimuli per day, regardless of how many hours a reviewer actually spends to complete the review of the assigned stimuli.

To summarize, our assumptions regarding the hiring and training of ELA stimuli accessibility reviewers are:

- Two reviewers per grade; 14 reviewers total
- Two hours of live, online webinar training per reviewer
- Reviewers work individually and will most likely log into the system from home during evening or weekend hours
- Each reviewer reviews half the stimuli for a grade span (approximately 42) over a period of a few weeks
- Each stimulus is reviewed by one reviewer; however, a reviewer may request that a second person review any given stimulus
- Average review rate of five stimuli per hour
- Reviewers are paid a flat daily rate based on 32 stimuli per day based on a 6.5 hour day

Item Accessibility Reviewers

Our proposal is based on the assumption that 28 accessibility reviewers will be needed. The reviewers may be organized either within or across content areas and grade levels. We propose that a single accessibility reviewer will review each item and that each reviewer will be routed workflows with a queue of items to review over several weeks. A reviewer may, however, request that an item be reviewed by a second reviewer. Such items will be routed to a second reviewer; we made the assumption that 10 percent of items will be routed to a second reviewer.

We determined the number of accessibility reviewers needed by making the assumptions listed below.

- Accessibility reviews occur in September and October 2012
- Reviewers will work individually and will most likely log into the system from home during evening or weekend hours
- Each reviewer will review items intermittently over several weeks for a total of 4 days (26 hours)
- Each reviewer will be contracted for a flat rate based on the daily item review rate plus two hours of training, based on a 6.5 hour day
- Items can be reviewed at the rate of 25 items per hour, including documenting the review response
- Each performance task, on average, will take 30 minutes to review and to document the review response
- Each item is reviewed by one reviewer; however, a reviewer may request that a second person review any given item

Table 11: Item Accessibility Review Participants Needed

		Number of items for 2012-13	Total items	Items to Review*	Review Rate per Person per Day	Total Number of Review Days	Number of Participants	Review days per Participant
ELA	Selected Response	1,183	6,090	6,699	163	41	14	4
	Technology-Enhanced Selected Response	1,680						
	Constructed Response	1,435						
	Technology-Enhanced Constructed Response	1,792						
	Performance Task	126	245	270	13	21		
	Technology-Enhanced Performance Task	119						
Math	Selected Response	1,183	5,957	6,553	163	40	14	4
	Technology-Enhanced Selected Response	2,030						
	Constructed Response	714						
	Technology-Enhanced	2,030						

		Number of items for 2012-13	Total items	Items to Review*	Review Rate per Person per Day	Total Number of Review Days	Number of Participants	Review days per Participant
	Constructed Response							
	Performance Task	126	245	270	13	21		
	Technology-Enhanced Performance Task	119						
TOTALS		12,537	12,537	13,792		123	28	

*Estimate that 10% of items receive a second review.

11. Obtain and track necessary copyright permissions for all relevant materials.

The Collaborative staff has extensive experience in selecting authentic material and in working with copyright holders to obtain permission for use of those materials. For all previously-published texts and stimuli in the public domain selected for items for the cognitive labs and small-scale trials, we will document the source and proper acknowledgment line in the database as part of the stimulus metadata. CTB will manage the process of clearing copyright permissions and updating the interim item authoring database with the copyright holder's contact information, terms of the use agreement, and any limitations on use.

Based on the responses to bidders' questions Q5, we understand that SBAC expects vendors to propose the length of time for the term of use for copyrighted materials. Based on our experience, we propose to do our best effort to negotiate agreements with rights holders for a minimum of five years.

Based on the response to Q49, we understand that payment of copyright fees is the responsibility of the vendor. However, please understand that copyright holders typically require payment of fees based on each time their materials are used in a publication, and the fees are based on the print or online quantity. While copyright holders typically do not grant perpetual permission for unknown or unlimited future use of materials, CTB will work with the rights holders to negotiate the most favorable re-use agreements possible (for a minimum of five years) so that materials may be used and paid for in future administrations. If rights holders will agree to granting unlimited use for at least a five year period for a reasonable fee that is within the range we have budgeted, we will certainly negotiate such reuse agreements.

Based on our extensive experience in negotiating copyright permission agreements and typical practice of paying fees for each published use, our price proposal is based on the assumption that payment of copyright fees for future, operational uses of the text or stimuli is not included. Future repeat uses are unknown and cannot be predicted at this time. Permission will be obtained for production of a limited number of Braille and Large-Print versions, as appropriate, for the 2012-2013 pilot administration. Copyright permission information, including the negotiated fees and any limitations on the permissions, will be stored in the approved database throughout the contract.

For all permissionable stimuli approved for use for developing items for the 2012-2013 pilots, CTB will secure and pay for copyright permissions for use of the materials in the 2012-2013 pilot administration.

12. Facilitate and manage content reviews, bias/sensitivity reviews, and accessibility reviews of all 2012–2013 pilot stimulus materials and items/tasks, using SBAC protocols and record keeping mechanisms.

The Collaborative has extensive experience in facilitating both stimulus and item reviews for complex, large-scale programs. Appendix E includes sample materials that we have used to conduct review meetings, including an agenda with workshop overview; training slides, including training in content and universal design and accessibility reviews; content, bias/sensitivity, and accessibility review checklists; protocol used to collect anecdotal feedback; and review metric summary spreadsheet.

The Collaborative will establish contracts with reviewers according to SBAC-approved protocols. There will be separate reviewers focused on content, bias and sensitivity, and accessibility. In order to maximize efficiency given the short time frame for developing the items for the 2012–2013 pilots, the Collaborative proposes that the three reviews be conducted simultaneously in addition to independently. Because reviewers will be located across the country, the Collaborative proposes that the reviews be conducted online, using secure protocols. The interim item authoring system that we are proposing to use will manage secure, electronic workflows whereby items are assigned to participants for review. The system will track the status of review workflows as well as reviewers' individual item ratings and recommendations or comments. Record-keeping will be conducted according SBAC-approved mechanisms, with the support of the documentation and tracking features of the item authoring system. Below is a description of how the reviews can be conducted.

Stimulus Reviews

- Reviewers are recruited proportionately from among the SBAC states, according to the participation policies and procedures (performed by the RFP 08 vendor).
- Separate reviewers are recruited for content review, accessibility review, and sensitivity/bias review. Content reviewers will be assigned according to their grade band of expertise.
- Reviewers are trained via webinar, for the specific purpose of their reviews, according to the protocols and requirements established through the work related to RFP 08.
- Stimuli will be routed to reviewers through workflows in the interim item authoring system. Along with each stimulus, a stimulus map can be routed, which outlines the salient points of each stimulus along with the standards that can be assessed via the stimulus. This would assist the reviewers in envisioning how the stimulus could be utilized.
- Reviewers will be given a due date for the completion of the review of the stimuli in the workflow and complete an online rating form/protocol for each stimulus. We anticipate that some review questions may require a yes/no answer, while others may use a rating scale. In addition, there may be a free-response comment section. There will be separate questionnaires for content, accessibility, and sensitivity/bias reviews.
- Responses to stimulus ratings are automatically recorded in the authoring system as the reviewer completes them. The results of the reviews will be compiled by the authoring system.
- Using SBAC-approved protocols, the facilitators of the review sessions will export a report of the review data and present the results to the SBAC-established final review committee to approve the stimuli to be developed.
- All review outcomes will be stored following SBAC-approved protocols and mechanisms.

Item/Task Reviews

- Selection of reviewers will follow the same SBAC-approved protocols as with the stimulus reviews (performed by the RFP 08 vendor).
- As with the stimulus reviews, separate reviewers are recruited for content review, accessibility review, and sensitivity/bias review. Content reviewers will be assigned according to their grade band of expertise.

- Reviewers are trained via webinar, according to the review segment, according to the protocols and requirements established through the work related to RFP 08. We anticipate that training will include:
    - Review guidelines and protocols
    - Overview of Item specifications and style guide criteria, as needed, to perform the reviews
    - item acceptability criteria
    - documentation protocols and response capture mechanisms
- The Collaborative proposes two rounds of reviews:
    - an initial, secure on-line review in which the reviewers are given a due date for the completion of the review of the items/tasks in the workflow and complete an online rating form/protocol for each and recommend to accept, reject, or revise the item/task. We anticipate that if they recommend to reject, a reason must be provided; if they recommend to revise, they must indicate the component needing revision.
    - a second review, in which an SBAC-approved panel of reviewers and the Collaborative content experts meet via webinar sessions to review and reconcile the committees' recommendations for revision and provide final approval on items/tasks to be piloted.
- All reviews will be documented and stored according to SBAC-approved protocols and mechanisms.

13. Revise items based on the results of reviews using SBAC protocols.

As described elsewhere, the interim item authoring system that we are proposing to use will track individual reviewers' ratings and comments on items. Prior to meeting with SBAC to review and reconcile the results of the reviews, CTB will pull reports from the system which consolidate and aggregate review ratings and recommendations for items and tasks, from the content, bias and sensitivity, and accessibility reviews.

Following the reviews, we recommend that items be triaged based on the following criteria:

- Items accepted for meeting most or all of the content, bias/sensitivity and accessibility criteria, including items recommended for acceptance with revisions
- Items meeting most criteria for two of the three types of review (content, bias/sensitivity, accessibility) but needing significant revisions for one of the types of reviews
- Items flagged for significant issues based on two or three of the reviews.

It is our recommendation that reviewers focus first on reconciling the items in the first two categories. Our item development plan includes an overage for both the vendor-written and educator-written items to allow for attrition during reviews and still yield 10,100 items and 420 performance tasks. For pricing purposes, we assumed that 80 percent of the items/tasks will have comments or recommendations for edits that need to be reviewed and, as appropriate, have edits made to the items. Based on responses to bidders' questions, particularly Q19 and Q24, our editors will incorporate edits based on the ratings and/or recommendations of the reviewers based on conformance with the SBAC item review guidelines, item and style specifications, and industry standards for quality items. We will flag items needing discussion for the review and reconciliation sessions between the editorial staff of the Collaborative and SBAC. We will determine with SBAC the threshold criteria for accepting or rejecting items based on the reviews (if such threshold criteria are not established as part of the work of RFP 08) or for performing triage on the items for final review and reconciliation of edits. These review sessions should include SBAC content experts, experts in bias/sensitivity review guidelines, and accessibility experts.

All reviews conducted by committee participants and between the Collaborative and SBAC will be documented and stored according to SBAC-approved protocols and mechanisms. The interim item authoring system keeps a version history of each item, and previous versions may be opened and examined.

Any item or task revision will be made following the same protocols/criteria used during item development and item reviews as described in SBAC 08. After revising an item or task, it will be essential to check the final version again to confirm that it meets all item review criteria for content, bias and sensitivity, and accessibility (i.e., to ensure that the revisions effectively and appropriately resolved the stated issues and do not introduce any new issues that would affect the validity, reliability, fairness, appropriateness, or accessibility of the item or task). All final versions of the revised items will go through a final QA review by CTB's trained Editorial Quality Assurance reviewers. This final review will focus exclusively on performing a final style editing and copy editing of the items. At this point, accepted items have been deemed to meet the requirements of the item specifications.

14. Coordinate with the SBAC psychometrics contractor to implement decision-making protocols to select items/tasks and stimulus materials for 2012-2013 pilots item/task pool.

CTB will ensure that selection for the 2012–2013 item/task pool is done in collaboration with the SBAC psychometrics contractor.

Members of the Collaborative team will work closely with the SBAC psychometric contractor to implement decision making protocols to select item/tasks and associated stimulus materials to construct the pool of items/tasks that will be the basis for the 2012/2013 pilot test administration. CTB staff routinely work with internal and external psychometric staff to construct item pools to support our large scale assessment programs. The protocols we develop with the SBAC psychometric contractor will result in a pool of items/tasks optimized to support the implementation of the SBAC assessments.

15. Coordinate and finance all face-to-face and online meetings and coordinated review processes (e.g., costs of meetings, travel and expenses, lodging and food).

CTB will be responsible for the logistics, facilities, and travel costs for all face-to-face and online meetings, and will coordinate all review processes. CTB will make provisions to conference in any Consortium staff members unable to travel to these meetings. Our Program Management staff will work closely with the SBAC to prepare and update any necessary materials or documentation prior to each meeting. We believe that the meetings and review sessions will be a vital part of effective program implementation, and as such, we will ensure that meeting outcomes are clearly understood, meeting agendas are distributed prior to the meeting, and meeting notes, including action items and decision logs, are distributed at the conclusion of the meeting.

16. Develop, implement and manage a detailed project and communication plan that will:

a. incorporate specific deliverables, milestones and incremental tasks for which the contractor will be responsible.

b. specify consortium members' responsibilities for applicable tasks.

c. identify specific review opportunities and associated schedule.

d. specify processes for version control and record maintenance.

e. identify other communication events.

Our management team will develop and maintain comprehensive project documentation and communications plans that ensure the transparency and oversight supports required for the delivery of all program deliverables. The primary forms of program documentation/communication include (each of these have been explained in greater detail in the Management Plan of this proposal):

- Program Work Plan - Top-level program planning document providing specifics relative to program scope and defined roles/responsibility.
- Program Master Schedule - Organized in a Work Breakdown Structure, the master schedule will provide the identification, organization and sequencing of all program tasks, deliverable and milestones.
- Deliverable Matrix - Definition of all major deliverables and assigned resources, delivery dates (including all dates for review cycles and approval).
- Status Reports - weekly management report on project status, key risks, upcoming events, critical required actions.
- Record of Decisions - Catalogued record for key decisions made during program activities
- Quarterly/Annual Program - Comprehensive program progress report

CTB will comply with all existing documentation formats as provided by SBAC in order to maintain consistency across SBAC projects and vendors.

17. Provide final report of work completed with documentation of records of communication and decision making for all aspects of work done.

CTB will complete all necessary program documentation in the form and formats required by SBAC. We have proposed a series of program documentation to ensure comprehensive documentation of records, communication and decision making for all aspects of program work (detailed in the Management Plan).

18. All SBAC documents and documentation should be provided in agreed-upon static and non-static formats

Understanding that SBAC has work underway with several vendors, we are most willing to propose formats and tools for project status reporting and documentation (and will have these ready for use), but are also most willing to adopt the documents and documentation formats already in use. Similarly, project collaboration tools and portals that are currently used to aid ongoing project communication will be readily adopted for use by our partnership.

A. Project Approach/Methodology

Part I Oversight, Item/Task/Stimulus Development and Reviews

Oversight: Coordination of all schedules and activities under the resulting contract(s) as well as ensuring communication related to the tasks within each part of the work and with other Contractors and SBAC leadership to accomplish all of the work.

CTB will maintain full responsibility for complete program oversight, including the maintenance of timelines and critical program deliverables, complete program communication across other contractors and SBAC leadership, and consolidation of all work completed on the contract. As detailed in this Proposal's Management Plan, we have assigned a very senior level management team with the skills and experience necessary to maintain program oversight and effectively coordinate all program activities.

The complexity and accelerated timeline of the program will require the team to build consensus quickly and move decisively in order to maintain the cadence required to develop, review, revise and deliver the items/tasks and stimulus within the required timelines. It will require a “one-department” mindset to develop quickly across the SBAC and our alliance organizations in order to deliver all that is required.

The team, under the leadership of the Senior Program Manager, will maintain a full program schedule and complete program documentation. A Master Program Schedule will be created to ensure identification, organization and sequencing of all project tasks, deliverables and milestones. The project schedule will take the key elements of the project and translate them into a time-based plan. The complete schedule will include a work breakdown structure, all tasks and activities associated with the project, and the interdependencies of the tasks to be performed. The Project Schedule will be created using CTB's project management scheduling software, and will be continuously monitored, updated and analyzed by the assigned Program Schedule Analyst.

The team will provide comprehensive program documentation using a number of management tools, including a Program Work Plan. This top-level program planning document provides the details of the program scope and defined roles and responsibilities of all participants to ensure everyone is able to be productive and effective right from the start. The Program Work Plan provides the answers to the who, what, when, and how, questions related to key activities, milestones, deliverables, timeline, resources, risks, program controls, and quality controls. This document will become the central control document for all teams working on the program, and the basis for change management. The work plans for each program part align to the overarching Program Work Plan, and will define at a greater level of detail the tasks, services and activities to accomplish the scope of each part.

Oversight: Recommend enhancements and improvements to SBAC guidelines and processes

The Collaborative recognizes that many things need to come together for the successful development of an item pool for the 2012–2013 Pilot Tests. Work under this RFP will need to integrate and implement multiple deliverables from other vendors. The Collaborative staff recognizes the challenges in combining item/task specifications, guidelines and training materials into an integrated whole to support item development. We will work closely with SBAC to develop a detailed implementation plan for using these documents.

As a first step, we will evaluate the training materials and training agendas. Given the innovative organization of editors and teachers that we are proposing, minor modifications to trading agendas may be needed to provide for multiple levels of training. We will begin evaluating training materials as soon as they are available and will present an implementation plan to SBAC as soon as practicable after contract award. At this time, we will also recommend any changes or modifications that we recommend prior to the beginning of training of item authors.

Training materials for the content, bias/sensitivity, and accessibility reviews will also be reviewed although these materials will be used later in the project. The information gathered during the implementation of the training materials for item authors will inform the implementation and recommended changes for the materials used for the committee reviews.

During training sessions, our facilitators and editors will take notes on how the training materials are being use by both facilitators and participants. These notes and annotations will be compiled into a final report for SBAC that describes how the materials were implemented and including recommendations for

- Enhancements to materials for facilitator use
- Enhancements to materials for participant use
- Improvements to training agendas
- Additional recommended materials/appendices

Our model for item authoring, describe later in the proposal, provides for ongoing feedback between editors and teacher authors for which the original training materials will be adapted. A final debriefing meeting with our editor/facilitators on the efficacy of the training materials will collect additional data for these reports.

L1-12 Write and Develop Pilot Test Stimulus Materials; Write Pilot Test Items and Tasks including Innovative Item/Task Types

Development of stimulus materials and items/tasks that represent a range of item/task/stimulus types aligned with SBAC content and item/task/stimulus specifications as well as items/tasks/stimuli that allow for measurement of content and aspects of learning that have historically been difficult to measure in large-scale assessment (e.g., synthesis, engagement with technology, and application of content) and/or items/tasks/stimulus materials that lead to increased measurement precision. Items and tasks will maximize measurement validity and reliability while minimizing testing system impact and burden on students. Technology enhancement of items/tasks/stimuli will span the gamut of item/task/stimulus types.

In this section, we present our detailed plan for item/task/stimulus development for the cognitive labs, small-scale trials, and 2012–2013 pilots. Throughout the development and review process, the Collaborative will focus on ensuring that all activities adhere to SBAC's procedures for participation by SBAC-member states' educators, training materials and processes, review guidelines and procedures, and feedback protocols. All activities will have the singular aim of developing a pool of test items and tasks that align with SBAC's vision for high-quality, innovative items/tasks/stimuli, and will adhere to the item specifications and style guide for creating CCSS-aligned assessments.

The Collaborative partners have decades of experience in directing and managing the development of test items by vendors, independent contractors, and educators. Appendix B provides samples of item specifications that CTB has used to train and guide item writers to develop items aligned with the content and item specifications. Appendix D provides samples of the item acceptance rubrics that CTB uses to evaluate the draft items submitted by writers. This appendix also includes an example of feedback comments given to writers, both to encourage them and to redirect their efforts. We will, of course, use SBAC materials, guidelines, and procedures for all aspects of item development and editorial reviews.

A Comprehensive Design and Development System

The development of stimuli, items, and performance tasks called for in this RFP is part of a comprehensive plan for the development of the SBAC balanced assessment system. The SBAC Content Specifications, Item/Task Specifications, and other documents are all grounded in the idea of evidence-centered design. This focus on evidence is a critical component of item development. CTB uses an assessment framework that is based on a comprehensive assessment design and development system; this framework will guide our implementation of the Item/Task Specifications, Guidelines, and Training Materials for the development of the item pool for the Pilot Tests. The eight steps in the system are, briefly:

1. Define intended inferences and uses of the assessment data.
2. Define the test construct(s) that will become assessment targets.
3. Develop initial proficiency level descriptors to guide interpretation of test scores.
4. Design the assessment blueprint and write item and task specifications that are consistent with the identified assessment targets and item types.
5. Develop items and performance tasks based on specifications.
6. Refine items and tasks through collaborative review by stakeholders.
7. Field test items and tasks in appropriate small- or large-scale settings.
8. Implement the operational test and continue the design and specification validation process.

The reverse engineering concept in this process is apparent in Step 1, where the intended inferences and uses are specified, based on SMARTER Balanced assessment scores. The assessment targets, performance expectations, and assessment blueprints in Steps 2-4 have been determined by SBAC. Step

5, develop items and tasks, is the focus of this RFP and response. All subsequent design steps, development, and validation efforts are intended to support and refine the inferences and uses in Steps 1-4. Evidence bases are explicitly required in all eight steps, including the fifth step, where items and tasks are developed. In implementing the activities of this proposal, we will facilitate the development of items and tasks within this comprehensive framework.

Our Innovative, Professional-Development Focused Approach

We understand how important it is to increase the assessment literacy skills of educators and to equip them with knowledge and skills that will help them help students to master the Common Core State Standards and to prepare them for college and career. We are proposing an approach to working with educators from SBAC states as item/task/stimuli writers that will foster their increased understanding of the CCSS and of the way to assess students' mastery of the content.

Given the scope of deliverables to be developed, including those that require innovative approaches, it is essential that a finely detailed development plan be executed. This plan must be aligned with assessment blueprints and consistent with item and assessment specifications, and must include the degree and type of measurement for all assessment targets. If awarded this contract, the Collaborative will work closely with the SBAC governing board to finalize this detailed plan and to execute it with efficiency and proficiency.

To successfully achieve this goal, the Collaborative proposes that the SBAC participant item/task writers be assigned to teams of three writers and one editor, based on their experience and skill sets. Based on the number of educator-written items for the 2012–2013 pilots, we propose between six and nine teams per grade band (3-5, 6-7, 8-HS) for both English Language Arts and Mathematics item development for a total of 21 teams for each content area—42 teams total. Each team will work with a single assessment editor from the Collaborative's partners. We are assigning 42 experienced assessment experts to this project, in addition to content development leads who will provide coordination and oversight. Each team of three writers will become “experts” in a specified area of item development, with each area based on the organization of the SBAC content specifications.

Each team of three participants will work with an editor. The editor will coach the team members in the use of the specific item specifications that the team will use to write items/tasks in accordance with the item development plan. The editor will schedule weekly online webinar sessions to discuss draft items as a team and to examine and analyze model items. The editor will be responsible for reviewing draft items submitted by the team members. The editor will provide feedback on individual items and direct the writers on how to revise items to more precisely align with the CCSS, adhere to the item specifications, and reflect the overall vision for the SBAC assessments.

Our plan is designed to develop high-quality items by writers’ becoming experts in a portion of the content and item specifications. By working with a set of item specifications, rather than a wide range of content and specifications, writers will have the opportunity to internalize and deeply understand the content and specifications. By working with team members in cross-grade teams whenever practical, writers will also develop an understanding of how the content builds across grades and how students' mastery of the content builds. Writers will be able to generalize their experience to other areas of the content and to other item specifications when they share and apply their knowledge and skills in their classrooms, schools, and districts. We believe that the model we propose is efficacious in developing and disseminating assessment literacy skills within the SBAC member states.

Additionally, because a team will be wholly responsible for the items aligned with the standards and item specifications assigned to them, the editor will know what items the team has written and will be able to direct the team to avoid repetition of ideas and near-clones of items. The editor will be able to suggest ideas for items to the writers, thus ensuring a wide variety of items that meet the requirements of the item specifications. The Collaborative’s editors will confer frequently with each other and with SBAC staff, to discuss and resolve issues as they arise. Editors, individually and collectively, will monitor the items produced to assure that rich and diverse collections of items are being developed, all of which reflect the vision and goal of the SBAC assessment.

Following the training of the writers using the SBAC training materials, the work plan for the small teams will be as follows:

- In an online team meeting, the editor facilitates a review of the team's item writing assignments and the content claims, standards, item formats, and item specifications related to the team's assignment.
- The team reviews the content standards and discusses how the standards progress across the grades and evidence that reveals students' mastery and progression toward college and career readiness.
- The editor presents sample model items (from the cognitive labs and small-scale trials) and guides the team through an analysis of the items, examining how the items demonstrate alignment with the CCSS and the item specifications.
- For the first assignment based on a mathematics standard or an ELA stimulus, the team discusses the overall approach and plan to writing a collection of items that fulfills the assignment, discussing specific ideas for items. Discussions will address technology-enhanced and technology-enabled items, and developing scoring rules for items.
- Item writers then work independently on their item writing assignments, submitting draft items to the editor. Writers may confer with their editor and other team members, as needed.
- The editor reviews the draft items and provides feedback to writers individually.
- The editor schedules weekly or bi-monthly brief team meetings to review progress, discuss items as a team, address issues, delve more deeply into the content standards and item specifications, and generate additional ideas for items.

At the beginning of each assignment, which is based on a claim and standard or an ELA stimulus, the team will repeat the series of brief online team meetings. As the team's work progresses and the team members gain greater proficiency, the team meetings should become briefer and less frequent. In order to facilitate the scheduling of team meetings, we will take into consideration the time zone in which participants live when forming teams.

In later sections we discuss our specific innovative plan for how ELA and mathematics item/task writers will be organized.

Contract with Stimulus, Item, Task Writers and Reviewers

Please refer to Objective and Scope of Work sections 7 through 9 for the discussion of our assumptions regarding the number of stimulus and item/task writers and reviewers needed and how payment will be determined. CTB will establish contracts with writers and reviewers, including confidentiality and non-disclosure agreements. The contracts will also establish that SBAC retains copyright and ownership of all content developed by the participants. CTB will work collaboratively with SBAC to develop the contract/agreement terms and language.

Training for Stimulus Developers/Writers and Item/Task Writers

Please refer to Objective and Scope of Work sections 7 through 10 and 12 for the discussion of how the Collaborative will use the SBAC training materials and protocols to train item/task/stimulus writers and reviewers. We propose that all training and reviews be conducted as online webinars. The online training webinars will be recorded and posted on the designated website for access and re-review by the participants. By having access to the recorded modules online, writers will be able to supplement their participation in the group training sessions with self-directed study. We will work closely with SBAC and the RFP 04 and 08 vendor, as appropriate, to assure that all training materials are used as intended and all training modules are delivered as designed. We will also use the quality control guidelines and procedures to obtain feedback from the trainers and participants regarding the training materials and procedures. We will use SBAC-established protocols and procedures to gather and submit feedback to SBAC.

Item Development Quantities

The item development tables presented earlier are repeated here for convenience and ease of reference. Upon contract award and the availability of the item specifications and blueprints developed by other vendors, we will collaborate with SBAC to develop the detailed item development plan which will specify the number of items by grade level, content area, claim, standard, item type, cognitive complexity, and other appropriate variables.

Our proposal is based on the assumption that all items/tasks/stimuli for the cognitive labs and the small-scale trials will be developed by the Collaborative. This is necessary, given the very brief time frame in which to complete this work. We have assumed that the distribution of items will be approximately in the same distribution as the items for the 2012–2013 pilots in terms of conventional or technology-enhanced and by item/task format.

For the items for the 2012–2013 pilots, we propose that SBAC educators will write approximately half of the items and tasks and that the Collaborative will write the other half. All items and tasks will be reviewed by committees of participants from SBAC member states.

Table 12: Cognitive Lab and Small-Scale Trials Item Development Plan

Content Area	Item Type	Cognitive Lab Items	Small-Scale Trials Items
ELA	Selected Response	49	309
	Technology-Enhanced Selected Response	64	442
	Constructed Response	56	379
	Technology-Enhanced Constructed Response	70	470
	Performance Task	7	12
	Technology-Enhanced Performance Task	14	20
	Stimuli	30	180
Math	Selected Response	49	321
	Technology-Enhanced Selected Response	78	545
	Constructed Response	28	189
	Technology-Enhanced Constructed Response	84	545
	Performance Task	7	12
	Technology-Enhanced Performance Task	14	20
Total Items		520	3264

Table 13: 2012-2013 Pilots Item Development Plan

		Vendor-Written Items			Educator-Written Items		
Content Area	**Item Type**	**Items Needed**	**Overage**	**Items to Write**	**Items Needed**	**Overage**	**Items to Write**
ELA	Selected Response	500	1.15	567	500	1.25	616
	Technology-Enhanced Selected Response	700	1.15	805	700	1.25	875
	Constructed Response	600	1.15	686	600	1.25	749
	Technology-Enhanced Constructed Response	750	1.15	861	750	1.25	931
	Performance Task	53	1.15	56	52	1.25	70
	Technology-Enhanced Performance Task	53	1.15	63	52	1.25	56
	Stimuli			295			295
Math	Selected Response	500	1.15	567	500	1.25	616
	Technology-Enhanced Selected Response	850	1.15	973	850	1.25	1057
	Constructed Response	300	1.15	343	300	1.25	371
	Technology-Enhanced Constructed Response	850	1.15	973	850	1.25	1057
	Performance Task	53	1.15	56	52	1.25	70
	Technology-Enhanced Performance Task	53	1.15	63	52	1.25	56
Total Items		5262		6013	5258		6524

In designing our proposed plan for stimulus, item, and task development, CTB's Chief Research Scientist, Dr. Wim van der Linden collaborated with our solution design team, guiding us in the development of our innovative approach to developing stimulus-based item sets. Dr. van der Linden leads CTB's research agenda, serves as a consultant on technical research issues to CTB research staff, and conducts research to advance the field of psychometrics. Dr. van der Linden is a global thought leader and authority in testing and measurement science, computer adaptive testing, automated test assembly, Item Response Theory (IRT), and the detection of cheating. His publications are considered seminal in the area of computer adaptive testing and automated test assembly, and have appeared in book form and in numerous peer reviewed and international journals. He is co-editor of three published volumes: Computerized Adaptive Testing: Theory and Applications (Boston: Kluwer, 2000; with C. A. W. Glas), and its sequel, Elements of Adaptive Testing (New York Springer, 2010; with C. A. W. Glas), and Handbook of Modern Item Response Theory (New York: Springer, 1997; with R. K. Hambleton). He is also the author of Linear Models for Optimal Test Design published by Springer (2005).

Because the SBAC assessments will be administered online using computer adaptive testing, the item pool for any given reading passage or language stimulus must be larger than the pool for a fixed form. We must also keep in mind that the passages and other stimuli must cover a wide range of cognitive complexity appropriate for the grade level. The items associated with each stimulus must also cover a wide range of content, difficulty, and depth of knowledge or cognitive complexity, because the CAT engine must have the flexibility that a relatively large pool of items offers from which to select items to present to students.

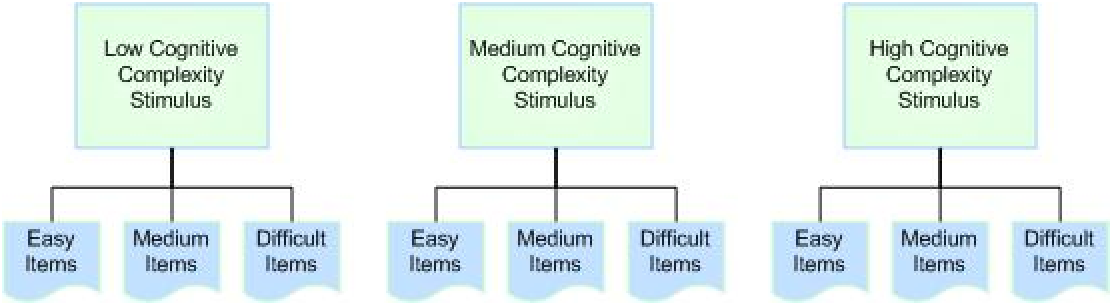
The need for a large pool of items (e.g., ~20-25 items per literary or informational text) introduces the challenge of avoiding items that clue the correct response to other items. On the other hand, a large pool of items allows for approaching the same concept in different ways. For example an item may ask students to identify the setting of a story, and another may ask students to identify a sentence from the story that reveals the setting. However, both items would not appear in the same test form.

We envision that ELA stimuli that meet the SBAC stimulus specifications might be conceptualized as being of low, medium, or high cognitive complexity. Tagging the complexity level of stimuli may be important so that the CAT engine selects a variety of stimuli for a student based on complexity in addition to content by claim, standard, and genre. This approach supports the assertion of ELA CCSS Reading standard 10 that students read texts with a range of complexity (CCR 10: "Read and comprehend complex literary and informational texts independently and proficiently.") We will work collaboratively with SBAC to define levels of cognitive complexity if such criteria are not included in the stimulus specifications.

We recommend that for any given passage, three testlets of items be developed. A testlet contains items that cover a range of content from across the claims, as appropriate to the stimulus, a range of cognitive complexity, and a range of any other attribute that becomes a constraint used by the CAT engine to select items. The difficulty level of items in a testlet, however, is targeted to a limited range of difficulty or performance level. Within a testlet, incidences of items that clue each other or that should not be presented together in the same test event, will be minimized. However, such "enemy items" may be in other testlets. Enemy items within and between testlets will be identified in the item database at the time of item authoring, and confirmed again when all reviews and edits are complete so that the CAT engine does not select an item that is an enemy of an item already presented to a student. This approach to creating testlets of items for a stimulus allows for the development of a wider range of items than what may be desirable for fixed form assessments.

The CAT engine, based on the estimate of the student's score to that point in the assessment, will select a text/stimulus based on cognitive complexity and then select a testlet of items from which to start selecting items to present to the student. Based on the student's response to items and the re-estimation of the student's score, the CAT engine may choose to shift to another testlet for the stimulus from which to select items to present. As an aside, the CAT engine can also be instructed to present a minimum and maximum number of items per stimulus to students in addition to a minimum and maximum number of stimuli by type. We would not want students to spend time reading a stimulus, only to be presented a few items for that stimulus.

Figure 1: CAT Testlet Development Plan



At the time of item development, the difficulty level of items will be estimated, guided by the information in the item specifications and the draft performance level descriptions. In order to develop the full range of items for a stimulus, the item set must be planned in advance as opposed to hoping that a writer will create such a set by happenstance. This necessitates that the stimulus be mapped and the items targeted for each testlet be planned as well. The final determination of the items in the testlets will be based on the pilot testing item parameters.

Our proposed plan to develop multiple testlets of items for each stimulus, in addition to providing the CAT engine flexibility in selecting items from a diverse pool of items, has the additional advantage of needing fewer stimuli. Of course, not all stimuli will be able to support 20 or more items. Poetry or stimuli for the lower elementary grades may not be lengthy or rich enough to support more than 12-15 items; however, other stimuli may certainly support more than 20.

Fewer stimuli translate into efficiencies—both in the development or selection of stimuli and in payment of copyright fees. Based on the response to bidders' question Q39, we understand that SBAC estimates 800 stimuli for ELA items for cognitive labs, small-scale trials, and the 2012-2013 pilots. While our price proposal is based on 800 stimuli (30 for cognitive labs, 180 for small-scale trials, and 590 for pilots), fewer stimuli are needed if, on average, 20 items per stimulus are targeted. We assume that 800 stimuli include an overage of stimuli to allow for attrition during SBAC or committee reviews which would not be replaced if stimuli are rejected. Based on the test blueprints and item specifications, we can work with SBAC to estimate a more precise number of stimuli needed in order to achieve efficiencies in stimulus development/selection.

Stimulus Development

As a first step, the Collaborative's ELA staff will become familiar with the stimulus specifications developed by the SBAC 04 vendor. The Collaborative's staff is already familiar with the ELA Common Core State Standards and the types and characteristics of stimuli that will be needed to assess the CCSS. We understand that innovative types of stimuli will be needed, including audio-video or other types of stimuli appropriate for technology-enhanced items. Audio-visual stimuli will present a particular challenge. Some types of video stimuli, such as of professional dramatic performances or film clips, will likely not be realistic or reasonable because of the expense and time to obtain copyright permission. Commissioning the creation of videos is also likely to be time and cost-prohibitive; however, commissioning the audio recording of a text in the public domain is certainly reasonable. Given these challenges, we will locate and pursue video and other stimuli sources in the public domain. CTB's copyright and permissions office staff is highly trained in copyright laws and experienced in clearing copyright permissions or investigating ownership to determine whether a stimulus actually is or is not in the public domain.

Our stimulus/item/task development plan assumes that 45 percent of the stimuli for the ELA assessments will be previously published texts, audios, or other stimuli, either requiring copyright permission or in the public domain. We anticipate that the percentage of commissioned texts and stimuli will be greater at the elementary grades while the percentage of previously-published texts and other stimuli will be greater at the secondary level.

We propose that the Collaborative will search for and select previously-published stimuli for the ELA assessments. We have extensive experience in selecting texts that are appropriate for assessing the CCSS as well as selecting texts from sources and rightsholders that have a proven track record of granting copyright permission for use of their materials in assessments or for online publishing. Also, our assessment and permissions specialists are well versed in what is or is not in the public domain. CTB has already identified and vetted approximately 70 texts in the public domain suitable for use for CCSS-aligned test items. These texts may be considered for the item development. All stimuli, regardless of source, will be of the high caliber and complexity required by the common core standards and the requirements of the SBAC stimulus specifications and necessary to produce the required variety and complexity of items and tasks.

Our proposal is based on the assumption that all stimulus-based ELA items developed for the cognitive labs and the small-scale trials will use either commissioned or public-domain texts or other stimuli. We have made this assumption because the use of the items beyond the cognitive labs and small-scale trials may include use as models, in training materials, as public release items, or in other non-secure ways. Based on the responses to bidders' questions, we understand that some items may be added to the pool of items for the 2012–2013 pilots. Using stimuli in the public domain or commissioned specifically for the SBAC item pool will allow the greatest flexibility in use without concern for the expense and time to clear copyright permissions for each use.

Copyright Permissions for Previously Published Materials

The Collaborative staff has extensive experience in selecting authentic material and in working with copyright holders to obtain permission for use of those materials. For all previously published texts and stimuli in the public domain selected for items for the cognitive labs and small-scale trials, we will document the source and proper acknowledgment line in the database as part of the stimulus metadata. CTB will manage the process of clearing copyright permissions and updating the interim item authoring database with the copyright holder's contact information, terms of the use agreement, and any limitations on use.

Based on the responses to bidders' questions Q5, we understand that SBAC expects vendors to propose the length of time for the term of use for copyrighted materials. Based on our experience, we propose to do our best effort to negotiate agreements with rights holders for a minimum of five years.

Based on the response to Q49, we understand that payment of copyright fees is the responsibility of the vendor. However, please understand that copyright holders typically require payment of fees based on each time their materials are used in a publication, and the fees are based on the print or online quantity. While copyright holders typically do not grant perpetual permission for unknown or unlimited future use of materials, CTB will work with the rights holders to negotiate the most favorable re-use agreements possible (for a minimum of five years) so that materials may be used and paid for in future administrations. If rights holders will agree to granting unlimited use for at least a five-year period for a reasonable fee that is within the range we have budgeted, we will certainly negotiate such reuse agreements.

Based on our extensive experience in negotiating copyright permission agreements and our typical practice of paying fees for each published use, our price proposal is based on the assumption that payment of copyright fees for future operational uses of the text or stimuli is not included. Future repeat uses are unknown and cannot be predicted at this time. Permission will be obtained for production of a limited number of Braille and large-print versions, as appropriate, for the 2012-2013 pilot administration. Copyright permission information, including the negotiated fees and any limitations on the permissions, will be stored in the approved database throughout the contract.

For all stimuli requiring permission and approved for use for developing items for the 2012-2013 pilots, CTB will secure and pay for copyright permissions for use of the materials in the 2012-2013 pilot administration.

Copyright Protection for Items/Tasks/Stimuli Obtained through Procurement Options

As part of the scope of work for Task 4, CTB will manage the process of contracting with the states and/or individuals who participate in the procurement option study. As part of the contracting process, CTB, with the support of the McGraw-Hill Education legal department, will develop agreements that participants must sign acknowledging that any items/tasks/stimuli developed for the study are works made for hire and are the property of SBAC. Similarly, for any previously-existing items contributed by states, CTB will work with SBAC to determine the terms under which SBAC will acquire and use the items.

Item Development and Review Process: Educator-Developed Passages and Items

The process of item development will be informed by feedback from other activities under the contract. To accommodate the tight timelines for item authoring and review, we are proposing phased development from early June through September. This phasing of item authoring will allow for an iterative item authoring process. In addition to inputs from Task 2 and Task 4, the Collaborative is proposing a value-added review by The College Board of a sample of grade 8 and HS items to give an indication of whether the final pool of items will meet the college- and career-readiness goals of SBAC.

Input to the item authoring process will come from three sources:

1. Research on automated scoring. Information from the research under Task 2 on automated scoring will be provided to item authors as it becomes available. Initially, we will be able to provide overall guidelines for authoring constructed-response items that can be scored using an automated scoring engine, and refine those guidelines as additional research results emerge. In Task 2, we propose beginning the scoring research with a sample of items from SBAC-member states. Using such a sample will allow early research findings to be incorporated into item development training, review, and editing processes throughout the authoring and reviewing window.
2. Data from cognitive labs and small-scale trials. The schedule we propose in Task 3 for the cognitive labs and small-scale trials will also allow research findings to be incorporated into the authoring and review process. The cognitive labs have been specifically designed to provide input to the development of technology-enhanced and other item types. Our proposed approach to the scope of work for Task 3 will provide important information to item developers about the effectiveness of various item formats and enhancements.
3. The College Board college- and career-readiness review. The College Board will review a sample of items at grade 8 and HS to determine whether they provide evidence of a student's readiness for college or a career as described under Task L1-19. The results of this review will be provided to those editors and authors working at grade 8 and HS to inform the development of items specifically at those levels.

These built-in feedback loops will support an iterative authoring and review process that will result in items that fully meet SBAC's definition of next-generation items and tasks.

Figure 2 below describes the Collaborative's proposed general process for educator-developed items, tasks, and stimuli, including the training of participants to develop the content and all processes related to reviewing items, tasks, and stimuli. Collectively, our team will provide 42 expert content staff members knowledgeable about the Common Core State Standards and large-scale item and assessment development, to serve as the small team editors. In addition to these 42 editors, we have assigned experienced staff to serve as overall content leads who will be the main point of contact between the Collaborative and SBAC on content matters. Our team will coordinate all logistics for the participant-developed item/task/stimuli development and review sessions, including the delivery of online training through webinars and the online review of participant-developed items. As requested, we have provided sample items created by the Collaborative partners in Appendix C. These samples provide evidence of the Collaborative's experience and expertise which we are excited to bring to the effort to create items/tasks/stimuli for the SBAC assessments.

We will also coordinate with the SBAC 08 vendor to communicate the number of writers needed for each grade and content area and to contract with the selected participants for writing approximately half of the items for the 2012–2013 pilots. We understand that participants for the content reviews, bias/sensitivity, and accessibility reviews will be recruited and selected by the RFP 08 vendor. We will contract with those educators who meet the selection criteria and are selected to serve as reviewer participants. We will provide performance and participation reports for each activity (e.g., item writing, item content review, and item bias/fairness, sensitivity review), which will be generated from the interim item authoring system.

Figure 2: Item/Task/Stimulus Development Overview

Step	Description
Item Development and Review Plan	Item development staff from the Collaborative, including editorial staff, will meet internally to review all guiding documentation (e.g., item specifications, sample items). Item development specialists and editorial specialists will generate a plan in accordance with the item work flow process charts. The plan will include a projected schedule for item development activities and all item development process steps including sign-offs, etc. A participant selection plan will also be developed for submission to the SBAC 08 vendor, including, but not limited to, identifying the number of participants needed by grade/content area and other areas of expertise, based on the approved item development plan. A high-level item development plan will identify the type and quantity of items each writer will be assigned to develop. Item development staff members from each company will meet with the SBAC leadership team to confirm our understanding of the program. The goal of this meeting will be to ensure that our team members have a clear understanding of the vision for item development, including the steps in the item development process, timelines, schedules, and the proposed number of items/tasks/stimuli to be developed by participants.
Writer Training	The Collaborative's staff will train participants on the use of the interim item authoring tool, as well as training them to write high-quality items to measure the standards. Training for each of the specific item types (e.g., selected-response, constructed-response, performance task, technology-enabled, and technology-enhanced items) will also be provided, using training materials developed by the SBAC 08 vendor. Training will ensure that item writers have an understanding of what makes a high-quality item in accordance with the SBAC vision for the assessment. The training will be conducted online. Our assumption is that the training will be provided as modules, beginning with a general overview of the SBAC assessment and becoming progressively more specific by grade and content area.
Item Authoring Tool (Item Bank)	Participants will write items and enter items directly into the interim item authoring tool (item bank). Item characteristic data stored in the system database will include, but may not be limited to, the following: alignment with standards, readability, cognitive complexity, conceptual description of graphic requirements, and estimated level of difficulty. Item descriptors (distractor analysis) will also be provided. Note: This process will occur throughout the item development process (for committee reviews).
Items Submitted for Review	After items are entered into the item authoring tool (item bank), CTB, DRC, CAE, and AIR item development specialists and editorial staff assigned to each small team of writers will review each item for acceptance or revision. Feedback on items will be provided to the writers, as appropriate, and items may be revised by the item writer, as needed.
Editing Cycles; Internal Item Review and Quality Processes	An editorial team from CTB, DRC, CAE and AIR will review items, tasks, and stimuli to ensure that all requested changes have been made and entered correctly. The College Board will review a sample of items for college- and career-readiness evidence.
Content Advisory Committee Review	Using the SBAC-approved training materials, training in content review will be provided. Committee members will review the items online. Items may be accepted as-is or recommended for revision. Some may be rejected. The status of all items and comments entered by reviewers will be captured online in the item authoring tool or item bank. (Note: feedback from the bias/fairness/sensitivity review will also be provided.)
Bias and Sensitivity Committee Review	Using the SBAC-approved training materials, training in bias/fairness/sensitivity will be provided. Committee members will review the items online. Items may be accepted as-is or recommended for revision. Some may be rejected. The status of all items and comments entered by reviewers will be captured online in the item authoring tool or item bank.

Step	Description
Accessibility Committee Review	Using the SBAC-approved training materials, training in accessibility review will be provided. Committee members will review the items online. Items may be accepted as-is or recommended for revision. Some may be rejected. The status of all items and comments entered by reviewers will be captured online in the item authoring tool or item bank.
Reports	CTB will pull reports from the interim item authoring system, compiling ratings and comments from the reviews. The final reports will reflect a summary of the item review statistics (e.g., number of items accepted without revision, number of items accepted with revision, number of items rejected, and comments).
Review and Reconciliation Sessions	Collaborative staff and SBAC staff will meet via online webinars to review items in accordance with SBAC criteria and protocols. The Collaborative staff will make final edits to items based on the reconciliation sessions. CTB's quality assurance editors will perform a final copy and style editing review.

Item Development Plan

Item Development Plan: English Language Arts

For the development of ELA items, we propose organizing educator item writers into a total of 21 teams. Each team will comprise three item writers and a dedicated editor, Each team will be responsible for developing selected-response (SR), constructed-response (CR), and technology-enhanced and technology-enabled (TE) SR and CR items and performance tasks for a specific grade and content area. Table 14 illustrates a set of teams for grade 3.

Table 14: Grade 3 ELA Teams

Team	Primary Area of Expertise	Item Types
3A	Reading Literary Text (Claim 1)	SR, CR, TE-SR, TE-CR
3B	Reading Informational Text (Claim 1)	SR, CR, TE-SR, TE-CR
3C	Writing (Claims 2 and 4)	PT, TE-PT

An experienced Collaborative assessment editor will lead each team, coordinate that team's assignments and tasks, and collaborate with editor leads of other teams. The editor will provide individual feedback to writers, and schedule regular team meetings that will be opportunities for planning item development for assignments and more in-depth coaching on producing items of each type aligned with the standards and item specifications.

Because the CCSS for ELA underscore the integrated nature of the content area’s concepts, skills, and processes, the item writing teams will also require some organizational flexibility as well as content facility. That is, while each ELA team may be assigned a specific grade span or an individual grade, text genre, assessment targets, and/or item types upon which to focus, it will be incumbent upon each writer to have an understanding of all item development for ELA. Thus, coordination and collaboration within the full ELA team and among the lead editors are vital to success.

The item writing assignments for each team will be based on the SBAC assessment blueprints and specifications. The Collaborative proposes that its editors will create the item writing assignments and, if requested, submit specific team-based plans (in addition to higher-level plans) to the SBAC governing board for approval.

While other approaches for each team’s focus have been considered (e.g., by item type alone or by standards), these alternate approaches are not practical in light of the integrative nature of the CCSS for ELA. Because texts are the key elements of an ELA standards/instruction/assessment cycle, it logically follows that they should be central to each team’s focus. Using Team 3A as illustration, this team will focus on literary texts, which will have been selected to support a range of literary genres, as appropriate for the grade, and with a range of text complexity. The team will write items to measure the determined assessment targets within the CCSS Strand Reading Literary Text. For each text, the team will develop SR and CR items—including those that are technology-enhanced or technology-enabled—that effectively address the depth and breadth of the texts' ideas, craft, and structure.

Additionally, because a percentage of texts will be paired so that students can be evaluated on their ability to integrate and synthesize ideas across texts, the Collaborative proposes, for efficiency, selecting texts that can be used as stand-alone stimuli or paired with another text. Most certainly, some of the paired texts will include both a literary and an informational text. For these pairings, coordination across teams is vital. Therefore, the Collaborative recommends that all texts be used first as stand-alone stimuli, and then re-distributed to the grade teams so that items addressing pairs of texts may then be written. For this purpose, it will be important that the Reading teams be knowledgeable about the specifications for both the literary and information for CCSS Reading strands and skilled in writing items for both.

The team assigned to Writing as a focus (e.g., 3C) will also use the texts selected for Reading measurement when the team is developing writing prompts and performance tasks. While the blueprint may call for some stand-alone writing prompts, it is most likely that Writing items and tasks will be in response to texts, and thus, coordination of the texts will be imperative for use by all teams.

Furthermore, it is logical that the Writing team also develops the performance tasks, for as noted in the SBAC's content specifications, "Full compositions, involving planning and revision, would best be assessed with performance tasks" (16. September 19, 2011 v19.0—2nd round DRAFT). It is further assumed that a percentage of the performance tasks will meet the requirements of ELA/Literacy Claim 4: "Students can engage appropriately in collaborative and independent inquiry to investigate/research topics, pose questions, and gather and present information." The Collaborative recommends that the writers for the Writing team be carefully selected, as their assignments will present the most challenges; these writers must be innovative and well-versed in all of the CCSS for ELA.

While not identified as a team focus, Speaking & Listening and Language Standards (Claims 3 and 5), will also be addressed per the requirements of the SBAC's blueprints and specifications. Depending on the skills being measured and the SBAC Claim into which each falls, Language items will be assigned to the Writing team or to a Reading team. For example, Language standards 4 and 5 may be measured with a text and, therefore, assigned to one of the Reading teams. Language standards 1 and 2 may be measured within students' compositions, with SR items associated with brief student-like written stimuli, or with discrete items; these items will be assigned to the Writing team.

The Collaborative will follow the SBAC's recommendations regarding how Listening standards are to be measured. If students will be listening to texts being read and then responding to SR or CR items, then the Reading teams will be assigned items to be written for that purpose, depending on the text's genre. We recommend that texts be selected specifically for use with Listening items because (1) texts for this purpose may have unique requirements and specifications, and (2) overuse of the same texts will result in limitations in the use of the text pool.

To support a CAT bank of items, it is essential that items be written that address a learning progression for an ELA concept or skill, and while every attempt to avoid clueing within a text's item set will be made, items that address similar ideas but at different points on the learning progression will be tagged as "enemies" within the pool of items for the stimulus. Students will not be presented these enemy items in one CAT assessment.

Item Development Plan: Mathematics

Just as for ELA development, we propose organizing mathematics writers and editors into 21 teams, each consisting of one editor and three writers. Each team will focus on a small area of mathematical content. However, the method used to determine which content should be assigned to each team will, of necessity, be somewhat different for mathematics than for ELA. This is due in part to the fact that mathematics items are discrete, rather than sets of items. In addition, the individual standards from the Common Core domains are scattered much more widely across the claims in mathematics than is the case for ELA. For these reasons, we propose assigning editors according to groupings of CCSS.

As an example, consider grade 8, which contains 28 assessment targets, spread across the four claims. The table below displays the number of assessment targets per Common Core domain.

Table 15: Grade 8 Assessment Targets

Domain	Assessment Targets
The Number System	2
Expressions and Equations	10
Functions	5
Geometry	8
Statistics and Probability	3

These assessment targets could then be divided as shown in the following table.

Table 16: Mathematics Grade 8 Teams

Team	Number of Targets Assigned	Domain(s) for Target Group
8A	8	Geometry
8B	10	Expressions and Equations
8C	10*	The Number System Functions Statistics and Probability

*total number across all three domains

Team 8A would have the responsibility to develop for all standards in the domain of Geometry that are found across the four claims. The same is true for Team 8B and the domain Expressions and Equations. Team 8C would handle the remainder of the assessment targets, of which there are 2-5 per domain. While the assessment targets are not evenly divided, this division is as equitable as possible without breaking up at least one domain among two teams. This same method could be used to determine the assignments for teams in other grades or grade spans (including high school), which would be dependent upon the grade-specific item development blueprints.

The rationale for having one team assigned to a relatively small area of content within a grade or grade span is that each team would then become expert in that area of content. All editors would understand the best way to apply the specifications and assess the individual skills in their purview for each type of item, and could then impart that knowledge to the writers, who would be able to apply their knowledge and expertise to the creation of all item types.

Other methods of assigning content were considered, including the use of learning progressions (LPs). Assignments based on LPs will help make certain that teams have a firm grasp of how assessment targets grow in complexity and cognitive demand from one grade to the next, thus making sure of a seamless transition from one year's summative test to the next within each LP. However, these issues will be addressed during the item specifications phase; that is where the best practices and limitations of assessing each skill within a claim must first be delineated. In addition, even across two grades, there will be a substantial number of items in any one LP (such as for Operations/Algebraic Thinking) based on SBAC claims; this would require more than one editor being assigned to that LP, which defeats the purpose of organizing editors according to LP. However, it will still be absolutely necessary, as described in the ELA section above, for teams to communicate and collaborate both within and across grades.

Collaboration among teams, both within and across grades, will be especially important during the development of performance tasks. It is recommended that performance tasks focus primarily or solely on one area of content, such as that found within a domain or a standard. However, it is common for performance tasks to also assess related skills found in other skills or even domains. Even though writers and editors will focus primarily on one small area of content, that does not mean that they will be unable to utilize other areas of the Common Core to create rich meaningful performance tasks. This is where blueprints, specifications, and the collaboration with other teams will prove most useful.

The members of the Collaborative have a great deal of experience in working with internal staff, vendors, consultants, and customers across the United States and internationally. We are all cognizant of the obstacles that are frequently encountered by this geographical spread. Our editors are flexible and adaptive to supporting item writers and responding to queries in a timely manner.

L1-18 Conduct Content, Accessibility, and Bias/Sensitivity Reviews of Stimulus Materials

> *Implement the content, accessibility and bias/sensitivity reviews of stimulus materials using the materials and procedures established in the work from RFP08; develop a pool of stimulus materials for the pilots; make recommendations about improvements to training materials and procedures.*

Please refer to the detailed information in Objective and Scope of Work sections 7 through 10 and 12 for how the Collaborative proposes to conduct the content, bias and sensitivity, and accessibility reviews of stimuli and items and tasks. These previous sections also discuss our proposed plans for collecting and aggregating review data and comments and compiling feedback for SBAC and conducting the final review and reconciliation sessions.

As part of the initial contract start-up meeting or soon thereafter, we will work collaboratively with SBAC to determine the protocols and mechanisms for collecting feedback on the content and usability of the training materials and item/task/stimuli specifications and how to report the feedback to SBAC. We propose the use of online surveys and the use of the quality control procedures developed by the SBAC 08 vendor. We will discuss options with SBAC and jointly determine the most efficient and efficacious means of gathering and reporting feedback.L1 – 19 Conduct Content, Accessibility and Bias/Sensitivity Pilot Item and Task Review Meetings

L1 – 19 Conduct Content, Accessibility and Bias/Sensitivity Pilot Item and Task Review Meetings

> *Implement the content, accessibility and bias/sensitivity reviews of items/tasks using the materials and procedures established in the work from RFP08 to select items/tasks for the pilots; make recommendations about improvements to training materials and procedures.*
>
> *Implement, evaluate, and respond with recommendations for all of the materials and processes (item/task/stimulus specifications, item writer training materials, content review, bias/sensitivity review, and accessibility review training materials and processes)*

The Collaborative will gather participant recommendations at specific process points throughout the entire stimulus/item/task stimulus and development cycle. We will use the quality control guidelines and procedures developed by the SBAC 08 contractor to evaluate the effectiveness of the online training modules and materials as well as the item/task/stimulus specifications.

We anticipate that participants' feedback and recommendations will be gathered using brief targeted online surveys of 8-12 questions. Surveys will be completed anonymously but will ask for certain relevant participant demographic information. We will want to differentiate among user variables (such as prior experience, etc.), training or document variables (such as clarity, completeness of training, etc.), and process variables (training time/schedule, workflow processes, etc.). A combination of variables may influence the usability and effectiveness of the specifications, materials, training, and processes, and it will be important to distinguish among them. The Collaborative will compile and analyze the results to produce a final set of recommendations and to measure the effectiveness of the process.

We recommended that participants' feedback on training materials or specifications be gathered at three points in the development or review process. We recommend this time sampling approach because we anticipate that participants' feedback may evolve over time as they gain experience with the use of item/task/stimulus specifications or with review guidelines and processes. As writers and reviewers delve deeper into the specifications and guidelines and internalize the information, their understanding of the information will evolve, and we would expect that the nature of their feedback may also evolve.

Item/Task Development Feedback

1. After item/task writing training, from both item writers and editors (to focus on clarity of item writing training and trainers, etc.)
2. After initial round of item batches, from both writers and editors (to focus on specifications and other writing materials, item review and revision process, effectiveness of editor feedback, etc.)
3. At the conclusion of the item writing phase (to focus on overall item development process, materials, schedule, etc.).

Content, Bias/Sensitivity, Accessibility Review Feedback

1. After item review training (to focus on the clarity and effectiveness of training materials and procedures, etc.)
2. At a midpoint during the review period after a few workflows of items have been reviewed
3. At the conclusion of the item reviews (to focus on overall review processes, guidelines, and effectiveness of training materials and modules, workflow procedures, etc.

In addition to gathering feedback from participants about materials and processes, the content lead and editorial staff of the Collaborative partners will provide feedback. We will gather feedback in both formal and informal ways. Formal methods may include surveys and focus groups. Informal methods may include annotations and comments that they mark on PDF copies of the item specifications, review guidelines and protocols, and training materials as they use them when working with item writers or reviewing and editing items. As noted elsewhere, the College Board will review a sample of grade 8 and high school items to determine whether the items will provide evidence toward determining students' college and career readiness. The College Board's review will provide another source of feedback on the item/task/stimulus specifications and item writer and review training materials.

L1 – 20 Revise Pilot Stimulus Materials, Items, and Tasks Based on Content, Accessibility, and Bias/Sensitivity Committee Feedback to Prepare them for 2012-2013 Pilot Tests.

Item/Task/Stimuli Final Review, Editing, and Approval

The Collaborative partners have a proven track record of developing valid and reliable test items. A critical component of the development process is to reconcile feedback from the content, bias and sensitivity, and accessibility reviewers and determine whether and how to revise the stimuli or items to improve the quality and accessibility. The proposed interim item authoring system will track reviewers' ratings and comments for each item. We will export reports compiling the item ratings and comments and prepare for the review and reconciliation meetings with SBAC. Our experienced editors will review the ratings and comments from the committees and will organize and prepare items, based on the recommended revisions, for the joint review and reconciliation meetings.

The Collaborative will conduct online webinar meetings with SBAC to review the results of the three reviews and reviewers' comments. We will determine how best to resolve any issues affecting content, bias and sensitivity, and/or accessibility. All item and task revisions will then be made in the interim item authoring system. For open-ended items (CR and ER) and performance tasks, revisions will also need to consider the impact on the corresponding scoring rubric. We will scrutinize each item and make certain that all aspects of the item have been reviewed.

Revised items will go through a final supervisory and an editorial quality review to assure that all comments from the content, bias/sensitivity, and accessibility reviews have been addressed.

CTB proposes a series of online review sessions of the edited items by a small committee of SBAC partners for final approval of the revised items. The final review should be done in light of the original review comments from the content, bias/sensitivity and accessibility review panels, and the approved item acceptability guidelines. At the final review, CTB proposes using an "accept or reject" category only. Reviewers would be asked to provide specific reasons for rejecting any revised items. We propose that these review and reconciliation webinar sessions between the Collaborative and SBAC occur over several weeks as items move through the review process in batches during September and October 2012.

Following the final editing, review, and approval of the items, the items and all associated stimuli, art, and metadata will be exported from the interim item authoring database and transferred to the SBAC 07 vendor for loading into the SBAC item authoring system.

Interim Item Authoring System and Delivery to SBAC Item/Task Authoring and Pooling System

We propose to use CTB's Item Authoring System (CIAS) as the interim online authoring system while the SBAC system is being developed and deployed. This section describes the capabilities of the CIAS in order to provide SBAC with the confidence that the CIAS has the capabilities and functionality to support the authoring, editing, and review of stimuli, items, and tasks. More importantly, the items and all the associated metadata will be able to be exported and transferred to the SBAC system. CTB's technology department will work collaboratively with the other vendor to map the XML tags of the interim system to the XML tags of the SBAC system and meet other interoperability requirements. CTB's Technology department will work closely with the SBAC 07 vendor to transfer the item database to the SBAC system. Following the loading of the content into the SBAC system, CTB will perform an audit on approximately 5 percent of conventional items and 15 percent of technology-enhanced items to ensure that the items and associated metadata transferred accurately.

Overview

CTB's methods and procedures and supporting software will be used to provide participants with the most effective tool for item authoring and review activities. The CTB item authoring system (CIAS) offers both a feature-rich secure Web-based item development platform for participants and an item management solution for administrators. The CIAS includes all the tools and easy-to-use technology needed to make item creation easy and effective:

- creation, review, and editing of item content through 100% Web-based authoring tools; no programming skills required
- importing of existing test content from external sources
- exporting item content in a variety of formats, including XML, PDF, and HTML, through the use of Application Program Interfaces (API) that the CIAS has available for migrating data in and out of external systems
- use of a variety of question types—from simple selected response items to performance tasks to interactive problem solving such as drag and drop
- easy-to-use what-you-see-is-what-you-get (WYSIWYG) visual editor that supports images, multimedia, and fully editable math equations (MathML Editor)
- version history of an item
- administrative tools to manage passwords, user roles, and user groups
- workflows to review, edit, and validate items for content, style, bias and sensitivity, and accessibility
- state-of-the-art computing hardware and application software.

Because content is created through an online Web interface, users can access the CIAS anytime from anywhere. Access control features such as unique passwords and usernames, make sure that only authorized individuals with direct responsibility for item development, review, and editing are allowed access to the system.

Figure 3: Convenient CAIS System Access

Item Authoring

The CIAS includes all the tools and technology to make online item creation easy and effective. Authorized users can create content based on new or existing item specifications, classify items to specific expectations, select from multiple item formats that are all tailored to developing assessment content, and align items with one or more Common Core State Standards.

Selected-response items of the multiple-choice type typically have four possible answer choices, including the correct answer and three distractors. Selected-response items also include true/false items. Constructed-response items, both short and extended, can be based on a single stimulus or on shared stimuli such as graphs, charts, reading passages, etc., and can include a series of questions that build on authentic "real world" scenarios ranging from simple to complex. Scoring of constructed-response items is performed based on rubrics that assign varying levels of credit for correct or partially correct answers. Student responses can be programmed for automated scoring mechanisms, in preparation for administration of items by the SBAC test delivery engine.

Authorized users can also use the CIAS for creating real-world scenario-based performance tasks that require students to synthesize, analyze, manage, and interact with information and ideas from a variety of content and sources, in order to demonstrate knowledge and higher-order thinking skills, including those that address college and career readiness and 21st century skills. In addition to the standard content authoring tools associated with item development to handle headers, stems, stimuli (including those shared by all or some of the items), text, art/images, scoring, etc., users can target each component item for editing separately. Furthermore, users can create performance tasks that either align as a whole with a single standard or in which each component aligns with a different standard.

The CIAS also allows users to create innovative test item types that use digital technology to improve measurement of content knowledge and essential skills that cannot be measured, or cannot be measured well, using traditional text-based multiple-choice items. The CTB authoring system provides an effective platform to build items that can tap higher levels of thinking and reasoning and depth of knowledge against the full range of the CCSS, while ensuring that the measurement is fair, accessible and inclusive of all student populations. The system takes advantage of sophisticated interactive online scenarios, such as those that use highlighting, drop-down menus, audio, video, and drag-and-drop, Data models and support for complex types of technology-enhanced items, such as hot spots and simulations, are being planned for future development. Users can manage these authoring processes using a combination of standard and specific tools for previewing, editing, and scoring. The CIAS automatically generates the XML code needed for rendering of the items/stimuli in a Flash or HTML5 environment.

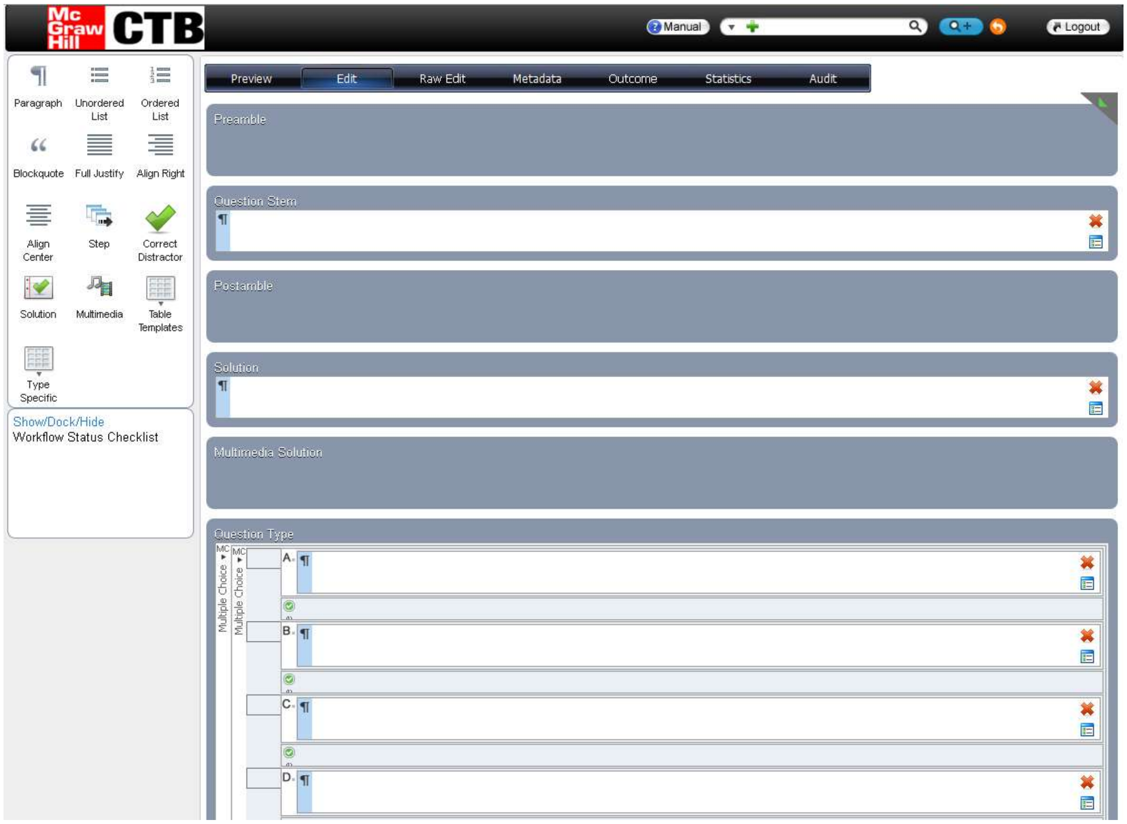
Anticipating that some users may not be familiar with the use of an online item authoring system, users with administrator rights can facilitate item authoring by having users automatically access pre-staged external templates/blueprints containing required item specifications or metadata.

The CIAS can easily be set up to accommodate expected item metadata, using numeric codes or brief labels. The metadata can vary with the type of item, passage, graph, and context of use. Metadata may include:

- content area
- grade/level
- item ID number
- common core state standard(s)
- item type
- content constraints
- cognitive complexity
- associated stimuli and rubrics
- scoring method
- Item/task author
- Item/task creation date
- details on the item or image(s)
- comments
- passage/stimulus length (i.e., word count)
- links to other stimulus materials
- copyright status and source,
- directions regarding formatting and style (symbols, text, notation, etc.) guidelines.

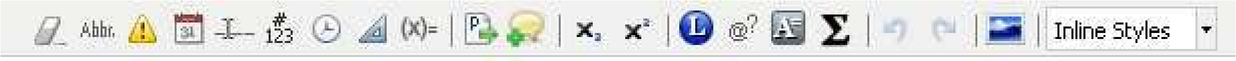
From the Item Edit screen of the CIAS (see Figure 4), users can employ a range of online item writing tools to enter new items or edit existing items, and review, save, and export finished fully approved items. The Item Edit Screen is divided into sections for easy entry of the information for each component of an item, such as the stem, the correct response and distractors, rationale for correct/incorrect responses, image properties, shared stimuli, multimedia, and other information that may be required to respond to the item or miscellaneous information to test the student's knowledge.

Figure 4: CAIS Item Edit Screen



For each section of the Item Edit Screen, a set of symbols allows for the customization of the layout and look of an item. Some symbols are specific to text editing, such as adjusting the alignment, bolding, italicizing, and underlining, adding/deleting text, and other functionalities.

Figure 5: CAIS Formatting Tools



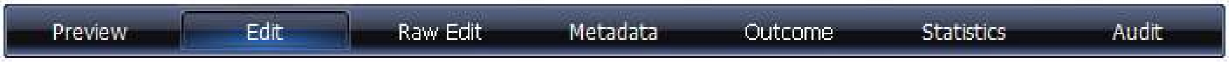
The CIAS can provide standard style directions for item writers in various ways. For example, if the style guide specifies that calendar dates must appear in test items formatted in a particular way—“YYYY/MM/DD’, there is a slide-out panel where the style guide can be placed for authors which would include the line "All dates must follow the YYYY/MM/DD format. Do not use any other format." Or, the CIAS can be configured so that upon saving an item, if it encounters the following string "2012/01/07" it can automatically tag it as a date. This means during export for publication transformations can be applied to ensure that the date does not break across lines, or it can be put out as "Saturday, January 7th, 2012." Additionally, if dates such as "August 4, 1945" or "9/11/2001" were entered, the system can be configured to convert those dates into the acceptable standard format OR produce an error requiring the user to make changes. Finally, if the style guide instructs writers to include both the symbolic and English-language versions of chemical symbols and formula (e.g., the text "NaCL" used in an item would also need to be tagged with the English "sodium chloride") to make the item more friendly for text-to-speech and/or Braille printing , the CTB system has a function to this by defining glossary items and attaching the glossary terms. The style rule can be made available to the user in the slide out panel as well.

Other symbols are specific to uploading images or creating mathematical symbols and complex expressions with MathML. By pressing the equation (Σ) symbol, the MathML Editor provides a flexibleuser-friendly graphic interface to facilitate the creation of equations and the selection of mathematical symbols within any piece of content. For example, there are several ways of entering MathML equations. Simple equations can be created using a standard keyboard (e.g., 2 + 4 = 6) by typing the equations as-is into the editor field. For more complex equations, the MathML Editor Shorthand tool translates the equation for the user. By using the "/" character, the MathML editor will automatically convert it into "—" for use in a proper fraction. For example:

2/3 would display as $\frac{2}{3}$

Finally, there are fully automated resource tools to preview and edit items, review item metadata, including in its XML format (raw edit), and audit logs to view history of changes to the item.

Figure 6: CAIS Resource Tools



Accessibility Features

The CIAS supports metadata for accessibility features that get applied at publication or runtime (i.e., when administered online). For example, it supports metadata for images that can be added to the publication export, which will allow for both Braille and text-to-speech. Braille is not a problem in most cases, but issues sometimes arise for Images and MathML. A number of projects worldwide to represent MathML in Braille and speech are being explored. The following scenarios describe the type of accessibility metadata that can be documented for an item developed in the CIAS with an image of a graph (in color) and some math equations and delivered in a Web-based application.

Student with a visual impairment: This student would employ standard accessibility tools like JAWS (a screen magnification program installed in Windows) to view and answer items. No additional information during authoring would be needed to make the item accessible, as the student can be entirely accommodated in the runtime environment. If the student needs more contrast than is provided in the color graphic, a black and white version of the graphic is automatically created and stored in the CIAS. The limiting factor here would be the runtime environment itself.

Blind student: This student would use a Braille reading keyboard to read the text, and some special software to convert the MathML to Braille. To make graphics accessible, every graphic needs an "alt-text" description, which is current functionality in the CIAS. An auditory description of each graphic could be recorded and added as sound file. This is not currently supported in the CIAS, but would be relatively easy to add. Finally, a text-to-speech processor external to the CIAS could be employed to convert the alt-text to an audio file and a batch process used to attach it to the image in the CIAS.

The alternative to this is using a text-to-speech processor to read the entire question. The text in the question would benefit from being additionally described with something like VoiceML markup. This is not currently supported in the CIAS, and a user interface would need to be designed to add this type of metadata.

Student with a moderate cognitive disability: This student might need a simplified version of an item. A second simplified version of the item would need to be authored. The CTB system has the capacity to store "alternate" versions of items. However an additional metadata field would need to be added to

categorize the alternate as the simplified version. The requirement to display these alternate items comes at runtime, where the testing system needs to know to ask for the alternate form.
As listed previously (above), the CTB authoring tool provides several metadata fields to capture copyright information. The metadata field can also be used to assure that copyrighted stimuli are used as their copyrights allow. Other fields can be added as custom metadata through the administrator interface.

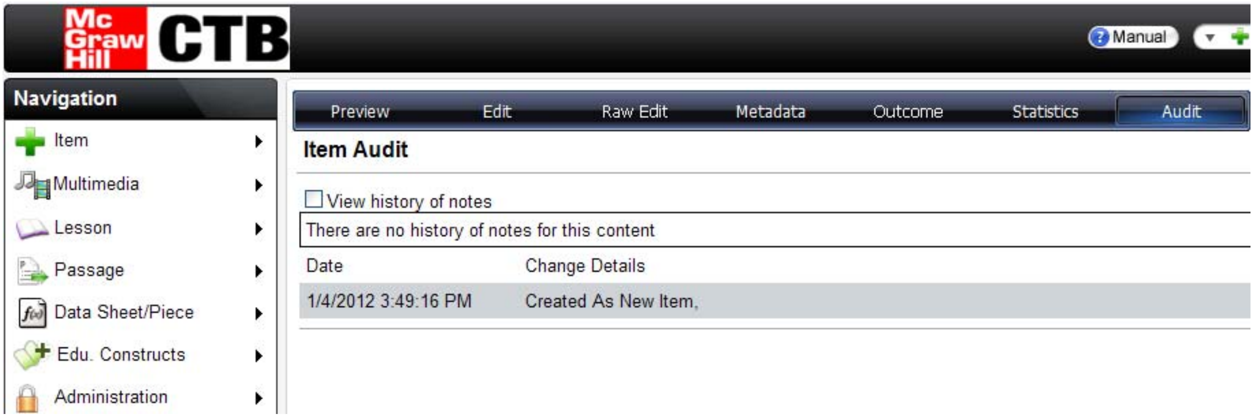
Content Management and Review
The CTB Item Authoring System allows for the use of workflows to manage content creation, and is capable of tracking the content by development status. The CIAS is capable of displaying the status of an item in any stage of the item life cycle. Furthermore, the system allows for the tracking of all changes to an item and reverting to a previous version if desired through the audit log. All of the workflows and item life cycle review processes are fully customizable to meet item acceptability criteria.

Authored items may travel through several workflows to facilitate content, style, bias and sensitivity, accessibility, and other types of reviews by authorized editors. Editors may return items to item writers with recommendations for additional work or approve, and submit them for transition to the next step in the item development cycle. Authored items that have completed a first review are placed in "Content Authored" status. Complete fully approved items are placed in "Ready for Export" status. Items can then be exported to other systems based on appropriate export criteria. Secure test content, including items under development, is transmitted electronically only by posting to secure FTP sites or through Simple Object Access Protocol (SOAP) Web Service.

All content is centrally located for easy access. Users with appropriate permissions are capable of searching for content based on a full text search or on content objectives/topics. The search engine is able to index any text-based information users define and then find it later based on various search criteria. The engine is flexible and able to search for multiple keywords using commas, "+" and "-" symbols. To search for a phrase, use quotation marks around the desired phrase to receive proper results. Users can also filter and order search results based on the desired attributes (item difficulty, item type, author, etc).

There is a basic messaging system built into the CIAS which is permission-based to record and track content, bias/sensitivity, and accessibility reviews. Also, through the use of permission-based workflows, one or more items can be assigned to one or more appropriate, authorized reviewers who can do their item review(s) online (asynchronous routing). The Item/User Audit report function provides a history of notes (see Figure 7). The User Tracking report function (see Figure 8) displays information related to authoring, editing, and reviewing that can be used to get performance metrics for every user or item.

Figure 7: Item/User Audit Reporting



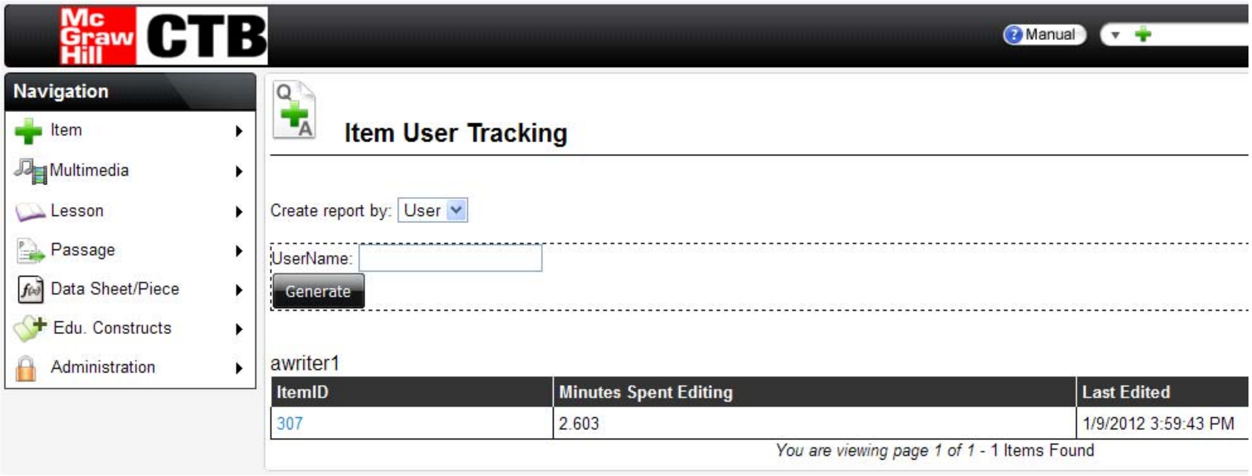
Item Audit

□ View history of notes

There are no history of notes for this content

Date	Change Details
1/4/2012 3:49:16 PM	Created As New Item.

Figure 8: Item User Tracking



Item User Tracking

Create report by: User

UserName:

awriter1

ItemID	Minutes Spent Editing	Last Edited
307	2.603	1/9/2012 3:59:43 PM

You are viewing page 1 of 1 - 1 Items Found

Administrative Functions

The CIAS offers a variety of different tools that are available to users with administrative rights. The tools can be used to create new users with the appropriate permissions, change passwords, manage user roles, create workflows, manage image properties, and perform various other functions.

Users may sometimes need to perform more than one function or will be asked to perform separate duties on a different group of items. In this case, their roles may need to be modified. For example, there is the possibility that a user has rights to write items but may also be assigned another user's items to perform a content edit. It is unnecessary to continuously switch the roles of such a user. An administration level user is able to manage the groups and to create a new group that has permissions to allow a user access to both areas.

The CTB system provides tools and reports to track item writing progress. Team Reports allow administrators see up to seven days' worth of data regarding any individual item or group of item writers, including how many items were created or moved within identified workflows for the period of time indicated, what changes were made, and who made the changes.

The use of workflows to manage the progression of work from start to finish ensures that items are assigned to the appropriate item writers, editors, and reviewers. The Workflow Wizard allows for items to be checked out and checked in, assuring that work is not being duplicated. The Workflow Wizard also allows the administrator to control and review workflow progress by status.

Interoperability

The CIAS has the capability to export content in various formats and through various mechanisms. In addition to PDF and HTML, the system generates complete item metadata in XML for export to other systems using an Application Program Interface (API) and Web Services interfaces.

In an effort to position the CIAS close to Accessible Portable Item Profile (APIP) compliance when the APIP standards are released (probably in mid-year), CTB has performed a thorough analysis of all of the APIP implications with regard to item authoring and online assessment delivery systems. CTB has begun the process of defining those XML tags that are expected to be required by APIP in order to determine those currently supported by the CIAS and those which will require further development to accommodate. Because CTB follows a very strict Universal Design protocol in all item development, the number of alternate accommodation renditions has been minimized.

As APIP becomes more clearly defined and accepted, CTB will continue to migrate its item content to include these new tags and to modify its export capabilities to provide the appropriate XML file formats for universal distribution. CTB can work with SBAC in the incorporation of the new APIP standards.

Security
The CIAS can be run securely over HTTPS. It has its own user account and group system. User Access control follows a permission-based model. Permissions are assigned to a group, and any users within that group are granted access to the permissions of that group. Some permissions control page-level access to features, while others control how much access a user has on a given page (read-only vs. write access, access to specific buttons or controls, etc.). Users can be members of multiple groups, and permissions for these groups are additive, i.e., a user will have all permissions from all of the groups the user is a member of.

Training
CTB's breadth of experience in item and test development has the necessary resources and tools to provide training to item writers and editors in the use of the system.

Part 2 Automated Scoring and Scoring Models

L1 – 15 Develop Comprehensive Scoring Approach

Introduction

The Collaborative is pleased to respond to the SBAC call to develop a comprehensive scoring approach to be tested as part of the 2012–13 Pilot Test. We propose a solution that will move the field forward in using automated scoring for items and tasks previously scored by hand. AIR will lead the Collaborative in developing the automated scoring approach, while leveraging the expertise of other Collaborative members.

Under this part of the contract, SBAC seeks to

- understand and extend the existing state of the art in automated scoring;
- identify, extend, and evaluate various open-source scoring engines; and
- on the basis of the abilities of automated scoring approaches demonstrated under this contract, develop a vision for scoring student responses from the pilot-test, field-test, and operational SBAC assessments.

The resulting vision will likely include the specification of automated scoring approaches, validation mechanisms, and integration with human scoring where necessary.

These are ambitious objectives, and SBAC hopes to achieve them on a very aggressive timeline. The CTB/AIR team offers a plan that will meet these timelines and objectives. The plan will

- support immediate work on the adaptation or development of open-source automated scoring solutions;
- make available for evaluation proprietary automated scoring solutions;
- allow the refinement and validation of those solutions based on data collected during the small-scale trials; and,
- optionally, help SBAC establish an organization and community of developers to support the future development and sustainability of the open-source solutions.

The timeline risks to this activity are critical. The small-scale trials constitute a key milestone on the critical path to develop and validate the scoring engines. Under Part 3 of this contract, we propose to begin cognitive labs immediately, but it is unrealistic to begin the small-scale trials (which will involve more than 30,000 students and 3,200 items and 64 performance tasks) before the beginning of the fall semester, 2012. This timeline leaves little time before the 2012-2013 pilot to revisit decisions made about the automated scoring engines. Therefore, beginning the evaluation of the scoring engines on this timeline poses unacceptable risk.

Instead, our team proposes begin the evaluation of automated scoring solutions almost immediately upon contract award. To support this schedule, we will begin with existing student responses, which we will solicit from SBAC member states. Success in this endeavor will require SBAC support. We recognize that many responses will have to be transcribed from paper, with the transcriptions preserving salient errors. We also recognize that the specific existing samples may not exactly match SBAC specs. However, finding the best existing examples from members will provide a test bed that will allow a timely start to this research.

Working with SBAC, AIR will identify four open-source solutions to serve as a basis. Each solution will be either an existing scoring engine or (for example) natural language processing tools possessing the features and capabilities needed for a scoring engine. AIR will

- enhance the open-source solution as necessary to work as a scoring engine;
- enhance the open-source solution to work within a standardized scoring framework to facilitate efficient experimentation, testing, and validation across the different engines; and
- place enhancements to the scoring engines into the public domain.

In addition to the open-source solutions, AIR offers some of our proprietary solutions for evaluation. In each case, we are willing to negotiate terms under which we can make these available under an open-source license, should SBAC choose to go forward with our proprietary solutions.

Finally, every example of durable sustained open-source software has had an organization behind it. In most cases, these organizations have had either a strong dedicated personality leading it or significant financial support from a corporation or an academic institution; in most cases, they have had both. SBAC's sustainability plan should establish a well-funded open-source organization to sustain the open-source software and nurture the community of developers that will be required to maintain and enhance it. We will be pleased to price as a separate option the establishment and staffing of an organization to begin this work.

Scope of Automated Scoring Engines

Assessment items may be selected response or constructed response. Selected-response items allow students to select from some number of choices, and the score on the item depends on these selections. Such items may be represented as multiple-choice, drag-and-drop, hotspot, hot-text, or other selection mechanisms. Such items, and the scoring rules for such items, fit neatly into the QTI framework. While multiple open-source QTI-compatible tools exist, the Java library, JQTI, may be the most mature and well developed. Where item rubrics are fully represented in QTI (selected-response items), they can be scored by the JQTI components that make up its scoring engine. JQTI is available under the open-source BSD license. AIR proposes using the JQTI library to evaluate rubrics expressed fully in QTI. We see the JTQI solution as a low-risk, open-source, QTI-compatible solution to the selected-response scoring engines, and we will use and evaluate that.

The core and challenge of this part of the contract is the scoring of constructed-response items. In the education field, all such engines have been referred to as artificial intelligence (AI) scoring engines, and the remainder of this write-up focuses on such scoring engines.

Background on Artificial Intelligence Scoring Approaches

Much of the history of artificial intelligence (AI) can be divided into two camps. In one camp are those who believe that intelligent systems should be modeled after our perceptions of human reasoning and, therefore, be transparent and able to "explain" their reasoning. In the other camp are those who assert that intelligence is rooted in quantum mechanics. Therefore, successful artificially intelligent artifacts are inherently inscrutable and cannot be based on traditional concepts of knowledge or logic. The first camp strives to build "glass box" artifacts or solutions, implying that the reasoning behind an artifact's actions is at least knowable. The second camp builds mostly "black box" artifacts and solutions, abandoning most attempts at systems whose actions can be explained.

Black box systems rely on statistical methods for building discrimination devices. They are very important for pattern recognition tasks. A number of developments in AI have evolved over the years, such as neural networks (aka associative networks), genetic algorithms, and simulated annealing. In automated scoring, Latent Semantic Analysis (LSA [1]) and Concept Indexing (CI [1]) represent statisticalblack-box approaches. Well-known scoring engines such as ETS's proprietary e-Rater system, AIR's Autoscore, and Pacific Metrics' TruScore software all use black-box approaches.

Black-box solutions built using statistical models usually require a training phase during which the artifact is subjected to exemplars that define the bounds within which it is to recognize patterns. Most often, the "reasoning" behind why such a system recognizes one pattern over another can be expressed mathematically, or described, but the explanations often make little sense from the perspective of the intended scoring rubrics. For example, LSA could explain assigning a score of x to auto-scored paper A by using a collection of words that was most similar to training paper B (or some cluster of training papers), which received score x.

Glass-box systems have relied primarily on various developments in knowledge representation. Classic examples include rule-based systems, semantic networks, taxonomies, formal grammars, automata, first- and second-order logic, and set theory. Artifacts built from these developments include expert systems, relational database management systems, programming languages, and compilers. In the realm of natural language processing (NLP), Princeton's Wordnet project is a fine example of a semantic network used to build a very large thesaurus for natural language. Systems designed and built using these techniques are most often handcrafted by humans.

In the field of automated scoring, AIR's Proposition Scorer and ETS's c-Rater represent proprietary glass-box solutions. These systems work by representing the correct answer, answers, or answer components as a series of concepts. Pattern-matching processes use transformational pseudo-grammars to search for the concepts expressed in the rubrics within student responses.

Some AI researchers have recognized that the black-box/glass-box dichotomy is counterproductive. Intelligent systems need both the low-level pattern recognition capability and the knowledge structures to make sense of the patterns and how they fit into a cognitive model of the world. We will show how a combined approach can benefit constructed-response item scoring.

Approaches to Automated Item Scoring

Even though structured representations of natural language have been studied for thousands of years, a sound (let alone complete) system has never emerged. Significant breakthroughs were made in the twentieth century, starting with Noam Chomsky's seminal paper "Aspects of the Theory of Syntax" [2]. Combined with advances in automata theory, artificial languages and parsing theory made possible the high-level computer programming languages that enabled the exponential growth in computing in everyday life.

As promising as these discoveries and inventions were, however, true computational linguistics as applied to natural language remained elusive and seemed intractable. This gave rise to a school of pursuit led by Roger Schank (dubbed Conceptual Dependency Theory [3]) that attempted to capture the concepts behind language. The intent was to permit quasi-grammatical structures to be understood by first converting them to an underlying canonical form of concept relationships that could then be compared with each other. True understanding was hampered by the growing realization of the vast amount of background knowledge that humans bring to the task of cognition as well as the seemingly infinite diversity capable within a natural language. (Even Chomsky's context-free grammar proves that an infinite number of utterances can be generated from a small finite set of grammar rules.)

The realization of the importance of background knowledge as a prerequisite for machine intelligence gave rise to two important projects: Princeton University's Wordnet and Cycorp's Cyc project. Wordnet is a machine-traversable network of relationships and equivalencies between words. Cyc is a vast digitized compendium of "common sense" knowledge organized in a way that permits inferences by automata.

The infinite diversity of language structures also gave rise to initiatives in applying statistical methods to language processing. Researchers began developing Markov models for part-of-speech tagging, word-sense disambiguation, and grammar discovery. Context-free grammars were "decorated" with probabilities on rule application. Probabilistic parsing accompanied them in an effort to tame the explosive nature of language ambiguity. This is a rich field of ongoing research today. The primary difference between statistical natural language processing and statistical essay scorers is that statistical NLP strives to use statistics and pattern recognition to tease out the meaning of texts. Statistical essay scorers are primarily interested in classifying the text features that improve correlation to human scorers whether or not any meaning is involved.

Pattern Recognition Methods

The inability of linguists to produce a reasonably strong model of language has given rise to statistical approaches to text feature recognition. Various pattern-matching algorithms are currently used for essay scoring. These generally rely on a training corpus that is representative of scored student responses to the assignment and are "annealed" until the desired correlation to human scorers is achieved.

The early projects (e.g., Project Essay Grade [4]) began from the premise that knowledge and structure of student texts are correlated with other more easily measured characteristics of the student text. Working from this premise, the systems used syntactic features of texts as proxies for deeper characteristics of the texts. Examples of such approximations are essay length as a proxy for depth of knowledge, word length as a proxy for sophistication of vocabulary, and counts of classes of words that signify complexity of sentence structure (such as prepositions and relative pronouns).

Advances in the art (e.g., Intelligent Essay Assessor (IEA) [5]) use LSA to analyze the training set by transforming word content into a two-dimensional matrix, with words forming the rows and documents (responses) forming the columns. Each row is treated as a dimension, and each document has a position in this highly dimensional space. Using standard matrix decomposition to pull out the principal components of this space reduces the dimensionality (much as factor analysis reduces the dimensionality of a set of survey questions). Subsequent operations take place in this reduced-dimensional space. In some LSA scoring applications, the score is assigned to match the nearest training paper. In others, the score is derived from the scores associated with the nearest cluster of papers.

In most commercial applications, including ETS's e-Rater and AIR's AutoScore, statistical techniques are married with natural language processing techniques to derive a score. Multiple features of the text, which can include LSA-like measures but also may include syntactic features and structural content features (such as the coherence of concepts within and across paragraphs), contribute to a statistical prediction of the score that a human would assign.

These approaches are black boxes. They do not endeavor to model the human processes that go into scoring them. Rather, they predict the outcomes of those processes.

Open-source essay graders exist, but available functional scoring engines do not appear to function at the level of commercial scorers. However, two open-source packages can provide a robust open-source basis for black-box scoring engines:

- The S-Space Package is a collection of algorithms for building Semantic Spaces as well as a highly scalable library for designing new distributional semantics algorithms. Distributional algorithms process text corpora and represent the semantic for words as high-dimensional feature vectors.

These approaches are known by many names, such as word spaces, semantic spaces, or distributed semantics, and rest on the Distributional Hypothesis: words that appear in similar contexts have similar meanings. The research and development is being done by the Natural Language Processing group at UCLA.

- Another package, SemanticVectors, creates semantic word space models from free natural language text. Such models are designed to represent words and documents in terms of underlying concepts. They can be used for many semantic (concept-aware) matching tasks, such as automatic thesaurus generation, knowledge representation, and concept matching. The concept mapping algorithms supported by the package include Random Projection, Latent Semantic Analysis (LSA), and Reflective Random Indexing. SemanticVectors was created by the University of Pittsburgh Office of Technology Management and is maintained by contributors from the University of Texas, Queensland University of Technology, the Austrian Research Institute for Artificial Intelligence, and Google, Inc.

AIR will collaborate with SBAC to decide which of these open-source packages provides the best basis for an open-source black-box scoring engine. Using the selected tool, we will build a black-box natural language scoring solution that is based on the capabilities of the selected tool. In addition, at SBAC request, we will include our Autoscore in the evaluation. If SBAC selects Autoscore, we are willing to negotiate an open-source license to use our engine.

Knowledge-Based Methods

A knowledge-based approach to item scoring requires the ability to create a model of the content knowledge required to demonstrate proficiency in the subject matter. Content models can take a wide variety of forms. Some approaches to simulations and automated tutoring systems attempt to build a broad deep representation of knowledge in a domain (Brown, 1978; Brown 1982; Burton, 1982) In some senses, the Cyc system mentioned above was an attempt to do just that for “common knowledge.” The Virtual Chemistry Lab (Yaron et al., 2010) represents a broad subset of the chemistry knowledge in a virtual laboratory simulation tool.

Other approaches begin with a broad but shallow knowledge base about a broad domain and then add deeper knowledge about a very shallow subset needed to score a particular task, item, or activity. AIR’s Proposition Scorer and ETS’s c-Rater both use this approach, with WordNet providing a shallow semantic lexicon and test developers building the concepts to be scored in student responses.

AIR’s Graphic Response Scoring Engine works on a similar principle. It has a built-in low-level knowledge of Cartesian planes. By referencing the low-level concepts, test developers can specify innumerable higher-level concepts, any of which may be used in the scoring of drawn lines, dragged objects, or other objects or shapes on the plane.

A true knowledge-based approach begins with a framework for the expression of many items within a defined content spectrum. A knowledge framework must consist of the following:

1. A theoretical foundation within the content domain(s) that defines and constrains the format of items
2. A language powerful enough to express all item responses within the domain above
3. A sound language for expressing machine rubrics
4. A computational model for applying machine rubrics to item responses.

If such a knowledge framework can be defined that is expansive enough to cover a large range of items within one or more content areas, then the item development time and effort are not only cost-effective, but also improve over time with experience.

Completing the example above, our graphic-response scoring engine satisfies each of these requirements. Its theoretical foundation is the Cartesian plane. Any item response that can be expressed in a Cartesian plane can be scored. This encompasses a very large domain of mathematics and other disciplines that use that domain. Graphic responses are captured in an underlying form of context-free language. Theoretical computer science provides the computational model for both recognizing and generating "sentences" in this language: the push-down automaton (Hopcroft, 1979).

The Graphic Response Scoring Engine's language for machine rubrics is first-order predicate calculus. For each score point, the item writer creates a statement of assertions in first-order logic that declares what must be true (and not true) of the response in order to receive the score. The computational model, a simplified resolution theorem prover, attempts to satisfy each rubric statement in descending order of score points.

Our *Proposition Scorer* for natural language responses is based on the same type of framework model as that for graphic responses. The main difference is in the way the atomic components of the assertions are teased out of the response. This item type is largely concerned with a student's ability to express in natural language the content knowledge that targets the item. The scoring engine requires a great deal of flexibility with respect to issues of spelling, synonyms, homonyms, and grammar. Various types of pre-processing cleanup, using resources such as Wordnet and non-deterministic parsers, is necessary in the attempt to provide the fairest score possible.

A third example of a knowledge-based glass-box system is an equation scorer. Our search of existing open-source materials found no equation scoring engines; however, many symbolic math processing engines are available, including Sage, SymPy, and Maxima, an open source version of the venerable MACSYMA [7]. AIR proposes to build an open-source, open-license equation scoring engine that tests a student response for equivalence to a correct answer provided by the test developer. This scoring engine will be based on one of the open-source symbolic math processors.

Our search of open-source materials has yielded no open-source glass-box alternatives for graphic or proposition scoring. As mentioned above, AIR is willing, at SBAC discretion, to evaluate these tools (or other commercial alternatives that SBAC may negotiate) for SBAC use. AIR is willing to negotiate the potential open-sourcing of these tools, as well as our AutoScore, if SBAC chooses to use them. In any event, if SBAC chooses to use these engines, and if an open-source solution is not negotiated, AIR will make them available to any other vendor under a commercial license, to ensure that robust and healthy competition ensues.

We note that glass-box solutions offer an inherent benefit that is not available from black-box solutions. Because the machine rubrics are explicit logical expressions of the content requirements of responses, it is straightforward to make them self-explanatory. By attaching natural language translations of the predicate logic for each score point, a knowledge-based scorer can provide greater insight into a student's strengths and weaknesses than can be obtained by a Likert-styled score at a subject or even a strand level. This extends the useful lifetime of secure summative assessment items. When items are retired and publicly released, they are often used as practice items on test preparation quizzes. If a constructed-response item previously used for a summative assessment can also provide a significant measure of diagnostic information, its usefulness is enhanced.

Brief Summary of Scoring Engines

As we describe below, we will develop or enhance five open-source scoring engines for constructed-response items. Table 17 provides a brief summary of the engines to be tested.

Table 17: Summary of scoring engines to be developed and evaluated

Item Type or Response Type	Open Source Solution	Proprietary Solution (optional)	Strategy to Ensure Continued Competition
Selected response, including drag and drop, hotspot, etc., as supported by QTI	Scoring engine based on JQTI	N/A	N/A
Essay or other long text	Up to two black-box engines based on S-Space or SemanticVectors	Autoscore; others that SBAC may negotiate to evaluate	AIR is willing to negotiate open-source licensing for our proprietary engines or offer commercial licenses to any other entity.
Other natural language-constructed response		Proposition Scorer; others that SBAC may negotiate to evaluate	
Graphic response	N/A	Graphic Response Item Scorer	
Equation response	Equivalence-matching engine based on open-source symbolic processor	N/A	N/A

Approach

Overview

AIR will develop up to four open-source, automated-scoring engines, and will evaluate up to six proprietary scoring engines, as described in Table 17.

As discussed in the introduction to Part 2, the first significant challenge that we will face will be timing. We propose to begin building the open-source engines immediately upon contract award. Having built similar proprietary engines provides us with insight into the actual time needed to develop the open-source versions. We anticipate that by bringing our experience to bear in this endeavor, we will be able to deploy the initial test versions of the engines very rapidly.

We envision the development of the engines to be an iterative process. We will develop the engines on the basis of existing research and technologies, and test them against student responses. By examining the cases in which the scoring engine fails to accurately score the papers, we will seek mechanisms to improve performance, modify the engine, retrain it, and reevaluate it.

To support this development, we will require a corpus of student responses on an existing set of items or tasks. We will solicit from SBAC member states. Success in this endeavor will require SBAC support. As described above, we recognize that many responses will have to be transcribed from paper.

By the time data from the small-scale trials are available, we expect that the scoring engines will have been through several rounds of revision. In most cases, we expect that the results from the trials will serve to validate and reinforce beliefs about the system.

At each iteration of the software, the testing of the engine will take place in two phases: a training phase and a validation phase. The validation phase will always take place using an independent sample. The scores from the engines will be evaluated on two dimensions:

- The ability of the engine to reproduce scores that match the most valid criterion scores available (typically, resolved human scores from a double-scoring process)
- The ability of the engine to detect and report cases that are unlikely to match.

The latter class of criteria is important as we help SBAC develop an overall approach for scoring. The overall approach for scoring will have to integrate in some way with human scoring. It is an often-forgotten fact that humans are involved even for scanning bubble-sheets to score multiple-choice items administered on paper. Sometimes scans are inconclusive, alignment marks are damaged, erasure criteria are not met, or other irregularities require that a human intervene in the process and examine the score.

It is simply unrealistic to believe that human writing will be automatically scored accurately for every case encountered. Some subset of responses will have to be scored or verified by human scorers. Scoring engines will enhance the efficiency of the system if they can identify the highest-risk papers accurately and automatically.

Work Plan

This section describes our plans to accomplish the work of this task. We have divided the work into three categories:

- Activities preceding the small-scale trials
- Activities following the trials
- Reporting activities.

Our initial activities, to be completed before the data from the small-scale trials are available, include gathering open-source software components and compiling a corpus of student responses on which to test the engines that emerge from the software work. Modification of the open-source components to score student responses will occur early in the project, as will the initial testing and iterative revisions to the software.

Prior to the trials, we will prepare the initial draft of the Vision Document, which will outline the needs and requirements of stakeholders. It will serve as the foundation for defining the requirements for the scoring system. Here, we use the term "scoring system" to refer to the set of automated and human processes by which final scores will be assigned to student work.

The next body of work begins as the data from the small-scale trials become available. From that activity must emerge a model for rangefinding and for refining the training of the scoring engines, including the materials themselves. We will conduct rangefinding sessions, hand-score and resolve the responses to prepare for training, validate automatically assigned scores, and conduct those tests on the scoring engines.

We discuss these activities below.

Activities Preceding Small-Scale Trials

Task 1: Gather and Prepare Existing Student Responses

An early start evaluating scoring approaches will reduce risk and increase the chances that SBAC will achieve robust open-source automated solutions. With this realization, we will ask SBAC and its member states to identify previous test items and tasks that approximate the types of items and tasks that SBAC would like to automatically score. We will ask SBAC leadership to assist in this effort.

AIR will work with individual states, drafting request letters and providing required confidentiality assurances. We expect that the endeavor will be greatly helped by SBAC leadership encouraging member states to participate.

We recognize that many of the items may have been administered on paper. These items will have to be transcribed onto the computer, preserving salient errors. We propose using a dual transcription process, in which one transcriber transcribes each paper and a second transcriber confirms the accuracy of the transcription. We distinguish salient errors, which should be preserved, from incidental errors, which should not. The best example of the latter is in capitalization. Many students will use capital letters in written text because they believe them to be more legible, not because they believe, for example, that all a’s should be capitalized.

We will work with SBAC representatives and, drawing on our experience with this sort of transcription, draft a set of transcription rules that leaves no room for transcriber judgment.

We believe that 10 to 20 prompts, with 500 to 1,000 student responses each, should suffice for this early training approach. For natural language scorers, the prompts should include a mix of short and extended constructed responses, prompt-based essays, and source-based essays.

Task 2: Gather Open-Source Components

Above, we outlined the best open-source components that we have found. For each candidate open-source component, AIR will summarize its capabilities and identify a list of benefits and liabilities of the tools. We submit this summary to SBAC to inform a discussion to select the tools to be integrated into the open-source scoring solutions.

AIR will establish a version control system (the open-source CVS) to contain these systems and modifications to them.

For each system, we will make the following modifications:

- Add the appropriate executive modules to apply the tool to automated scoring
- Integrate the engine with a common scoring interface to facilitate experimentation and testing.

For each open-source scorer to be developed, AIR will begin with a requirements document. The requirements document will identify the types of response to which SBAC hopes to successfully apply the scoring engine. The definition of success and the realization of these hopes are not guaranteed by this contract because SBAC aspires to extend beyond the state of the art. The requirements document will spell out details that include such aspirations as the time required to process each response, training requirements, the types of training samples that the system should accept, the handling of non-standard text (bad grammar, misspellings, etc.), and other features of the system.

AIR will prepare and deliver a detailed design document and unit test plan for the scorers, a plan that we intend to develop. End-to-end testing will, of course, entail the validation of the application of the engine. This process will have two types of outputs. It will certainly report the success of the engines with various types of input. In addition, we will investigate and report the types of items or responses with which the automated scoring is successful.

As described above, the “testing” (discussed in greater detail below) will also evaluate the ability of the engine to self-report high-risk cases—those that may not be scored correctly.

Task 3: Conduct Initial Analysis

With the first versions of the scoring engines in hand we will begin to test them against the sample of items and responses gathered from the states. Each engine will be trained (or, in the case of knowledge-based systems, rubrics will be defined).

We will divide the student responses into two groups: one to use for training and a second independent set to use for validation. The division of the items will be random; however, we will use a stratified random sample of responses. Often, few papers receive the lowest or highest possible score points. Obtaining a useful sample of extreme-score papers often requires having a very large random sample from the population. An alternative is to stratify the training set by the human-assigned score and then sample randomly within each stratum. In this case, the engines will have to accommodate unequal weighting of cases to accurately reflect the expected distribution of responses in the population.

AIR will work with SBAC to identify the outcome measures on which to judge success. In each case, we will compare the success of the scoring engines with human-scored counterparts. Likely candidates for the statistics to measure the correspondence between human scores and machine scores follow:

- Percentage of cases that match the (resolved final) human score exactly (exact match)

- Percentage of cases that match the resolved final human score within 1 point (adjacent match)
- Exact and adjacent match by score point
- Polychoric correlation between the final human score and the machine-assigned score.

These measures indicate the correspondence between the human scores and the machine scores. Each of these statistics will be examined for the overall population and for students in different subgroups, if demographic information is available.

Where we see mismatches between human-assigned scores and machine-assigned scores, we will examine these papers to glean insight into potential enhancements to the scoring approach.

As described above, it is also desirable for a scoring engine to identify cases that it cannot accurately score. As with any such endeavor, there will necessarily be a tradeoff between "false positives," the cases identified as wrong that are actually scored correctly, and "false negatives," incorrectly scored cases that are not identified as such. These possibilities are summarized in Table 18.

Table 18: Summary of possible outcomes in scoring accuracy classification analysis:

Flagged as Likely Scored Incorrectly	Actually Scored Correctly	Actually Scored Incorrectly
No	A	B
Yes	C	D

In this classification analysis, a scoring engine that perfectly identified the cases that it could not score would place all cases in cells A and C. (An ideal engine would place all cases in cell A.) Perfect prediction, however, is unlikely.

If false positives and false negatives are each equally detrimental, the objective would be to minimize the ratio $\frac{B+C}{A+B+C+D}$. However, false negatives (B) may carry more risk for SBAC. In this case, the index could weight those cases more heavily. AIR will work with SBAC to identify the specifics of the measures to be used to evaluate each engine's self-diagnosis.

Task 4: Prepare Vision Document

Early in the process, AIR will work with SBAC to develop the Vision Document outlining the vision for the scoring system. A vision document is a type of project charter. It defines the purpose and scope of the project before the detailed requirements document is drafted.

The Vision Document will express the goals of a comprehensive scoring system. These goals will be related to the current state of the art, and the vision will set forth a realistic path from current capabilities to future aspirations. Definition of the current state of the art will reflect the combined industry experience of the CTB-led team, a review of the literatures, and initial findings from the preliminary analysis.

Although vision documents grew up in the software industry, this project goes beyond software. The goal is to meet the scoring needs of the SBAC consortium and will necessarily include some hybrid integration of human scoring systems (which may be distributed and teacher-staffed) and automated scoring systems. It will include ongoing validations and, most likely, some sort of ongoing validity audit.

AIR will work with SBAC representatives to define this vision. The Vision Document will be concise, so that it may efficiently serve as a touchstone for decisions made during the requirements analysis and definition process.

Task 5: Prepare Recommendations for Validation of Scoring Models

This task seeks to "validate" the machine-scores assigned to items. Validity, however, is a characteristic of statements about or interpretations of test scores, not scores directly. Demonstrating that machine-assigned scores are comparable to human-assigned scores (particularly resolved human scores) provides validity evidence supporting the belief that the machine assigns scores as humans do. As discussed above, we intend to do such analysis.

Under SBAC's theory of action, scores on items reflect characteristics of student performances that provide evidence for claims about what students know and can do. To validate these claims we should seek evidence to support the claim that students with higher scores are more likely to have the claimed skills and knowledge and that students with lower scores on an item are less likely to have the claimed skills. More specifically, our purpose here is to evaluate whether the scores assigned by automated scoring engines provide any less (or more) evidence that students have the claimed skills than other evidence from the assessment.

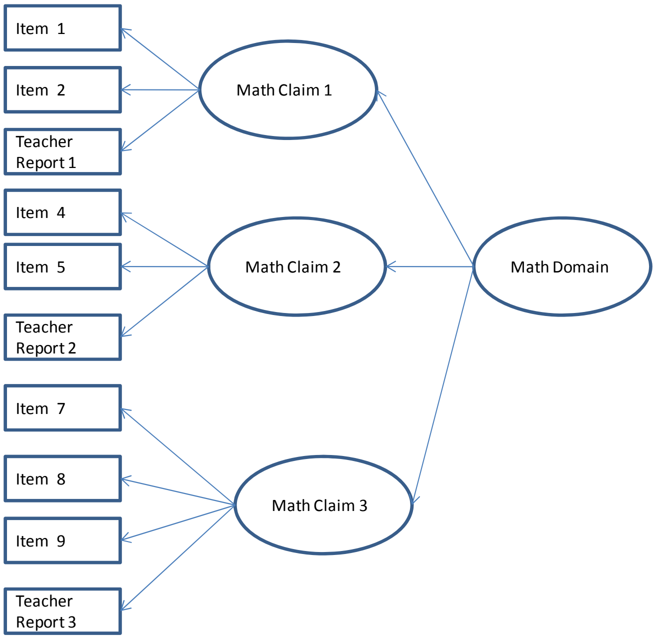
Unfortunately, we do not have a gold standard of measurement for each claim; indeed, if we knew clearly which students had which skills, an assessment would be unnecessary. The best available approach to evaluating the effect of automated scoring on the validity of the claims is to rely on multiple measures of multiple traits (claims). The assessment itself will provide multiple sources of information about each claim measured for each student. Some evidence will come from selected-response items, others from constructed-response items; still other evidence will come from technology-enhanced items. The design for the small-scale trials presented in Part 3 of this proposal ensures the availability of multiple measures providing evidence of each claim.

In Part 3 of this proposal we also propose to collect additional validity evidence for these claims. Specifically, within each grade and subject, we will ask the classroom teachers instructing participating students to rate each student on each claim. This will provide an additional independent measure of the student's claimed knowledge and skills.

Within the framework of a structural equation model, we can evaluate whether the evidence supports the inference that the machine-assigned (or human-assigned) scores support the intended claims. Consider the schematic presented in Figure 9 for example. We would expect very high loadings of the secondary latent traits (the claims) on the primary traits. The items intended to measure those claims would load on those claims, and substantial loadings would suggest that the items provide evidence for the claims. Moreover, this approach allows us to differentiate limitations of claims due to items from limitations due to machine- versus human-assigned scores. Items that show consistently low loadings for claims across scoring approaches provide poor evidence of the intended claim. However, items that show differential loadings across scoring procedures indicate that the quality of evidence for the intended claim is being compromised by a particular scoring procedure.

Responses to the constructed-response items could be tested with human and automatically assigned scores to detect systematic differences. The model could be extended to recognize the existence of rater effects in the teacher-reported ratings, improving the precision of the model.

Figure 9: Schematic illustration of potential analysis



Activities Following Small-Scale Trials

Task 6: Develop a Model for Rangefinding and for Training the Engine

Data from the small-scale trial will be used in rangefinding and in training and validating the scoring engines. Traditional rangefinding sessions take place in person. With the geographic distribution of the SBAC states, a more distributed model will prove more cost-effective.

AIR has used online displays of student responses successfully in rangefinding sessions. Rather than review a stack of papers, committee members sit together and view a screen on which student responses are displayed. This process has been used in multiple states for multiple years. In addition to reducing printing costs, logistical headaches, and security concerns, it allows committee members to view student online interactive work in a more realistic way than a printout would.

We propose to conduct rangefinding using this method under two different conditions:

- Under the first condition, committee members will meet in person and view student responses on the screen.
- Under the second condition, rangefinding will be conducted remotely using WebEx.

We will compare the two approaches based on:

- participant feedback on a survey form;
- SBAC-perceived quality of committee member participation during the meetings;
- SBAC-perceived quality of the resulting rangefinding material; and
- reported completeness and quality of the resulting training materials from the scoring partner (DRC).

Materials from rangefinding will feed into the scoring process, forming the foundation for human scoring. Depending on the final automated scoring approach taken and the nature of the items, the results of the rangefinding may feed directly into the scoring engines.

In any event, validation of the scoring engines will require that the constructed-response items be scored by human scorers. To obtain the most valid human scores possible, we intend to double score each response and have an expert scorer resolve all discrepancies. The scoring processes are discussed in more detail under Task 3.

Task 7: Materials and Processes for Rangefinding and Engine Training

Using student responses from the small-scale trials, AIR will develop a process for selecting responses for rangefinding and for training the engine. As discussed under Task 6, we will test two different models for rangefinding—traditional in-person meetings and distributed rangefinding conducted over the Internet.

Under both models, AIR will select the student responses to bring to rangefinding. Responses will be selected to identify the full range of score points, as well as typical and atypical responses at each score point. These approaches may be automated, may be implemented by human scorers, or may use a hybrid approach.

In all likelihood, different selection mechanisms will be used for different types of items. For example, items that are intended to be scored using a glass-box approach will have scores. In the past, AIR has used multiple-choice items on the same test to identify high- and low-performing students. When selecting rangefinding papers, we disproportionately sample high scores on the target items received by otherwise low-scoring students and low scores on the target items received by otherwise high-scoring students.

Black-box approaches do not support this sort of selection. The automated scores do not become available until after rangefinding, hand scoring, and training of the engine. In these cases, we often select a slightly larger sample of responses, stratifying by students' performance on the machine-scored items on the test. Using this approach, we disproportionately sample from high- and low-performing students, which increases the likelihood of finding responses at the extreme score points. The samples are then reviewed by AIR content experts and CTB/DRC scoring experts to select the papers that will prove most informative to the rangefinding committee.

Task 8: Conduct Rangefinding

AIR will conduct rangefinding by using the response samples selected under Task 7, implementing the models described under Task 6. Each rangefinding session, on line or in person, will be attended by an AIR content expert and a DRC scoring expert.

AIR will work with SBAC leadership and state representatives to identify potential participants for the rangefinding committees. We propose that each committee consist of five to seven educators, and that they each meet for two days to evaluate 36 constructed-response items for a single grade and subject. We will require a total of 14 committees to evaluate the full set of constructed-response items.

AIR will contact and recruit candidate participants. We will handle all logistics and costs associated with travel, stipends at the rate of $150/day, per diem, and other costs associated with the rangefinding meetings.

Task 9: Conduct Hand Scoring

Under Part 3 of this proposal we describe the details of our approach to hand scoring the responses from the trials. Because these scores will be used to train and validate the engines, each response will be double scored and every discrepancy will be resolved. Under Part 3 we also describe how we will use this opportunity to test different approaches to scoring items and tasks.

Task 10: Train Engine and Evaluate Success, Validate Against Human Scoring for Different Types of Prompts

Under this task, AIR will train the engines and repeat the analysis described under Tasks 2 and 3.

Using data from the small-scale trials on items and tasks written to SBAC specifications, we will train the engine on a subsample of the available responses. As we discuss under Task 3, 150 to 200 responses are likely too few to support both the training and validation of the scoring engines. Therefore, we have targeted a sample of up to 400 responses for each constructed-response item. We will use approximately 250 of these responses to train the engine, reserving a randomly selected 150 to validate the resulting scoring mechanism.

We will evaluate the success of these trials by using the same analyses developed under Task 3. These analyses will be conducted separately for different types of constructed-response items. These analyses will identify the scoring approaches that are successful for each type of item.

Reporting

AIR will prepare a report documenting the activities under this part of the contract. We envision the report to be iterative, with AIR delivering a draft to SBAC, revising the report on the basis of SBAC comments, delivering the revised report to the Technical Advisory Committee, and making final revisions in response to their comments.

The final report will address two main topics: rangefinding and scoring approaches. In discussing rangefinding, the report will document the models of rangefinding tested during this phase of the project and summarize findings and recommendations. The report will document the scoring engines developed and tested, and provide recommendation about the types of items and tasks to which each may be applicable. The report will describe how automated scoring might be augmented with human scoring to increase the validity of the scores.

Figure 10 provides a tentative outline for the final report. AIR will work with SBAC to refine this tentative outline to ensure that the final report meets SBAC needs. The recommendations presented in the final report will inform the processes to be tested in the pilot test in 2013.

Figure 10: Tentative outline for final report

1. Introduction: Vision for Comprehensive Scoring Approach
    a. Summary of vision document, as revised over the course of the study
    b. Summary of the organization of the report
2. Recommendations for Rangefinding
    a. Introduction: Summary recommendations
    b. Study design: Summary of models of rangefinding considered and measures of success
    c. Findings: Presentation of key data and findings from the testing of different rangefinding models
    d. Recommendations: Detailed recommendations about rangefinding
3. Recommendations for Scoring Approaches
    a. Introduction: Summary of recommendations
    b. Background
        i. Summary of the state of the art in automated scoring in each are enumerated in Table **<XX>**
        ii. Engines evaluated: Description of scoring engines evaluated
    c. Evaluation of scoring engines
        i. Design of the evaluation of the scoring approaches, including training, validation, and metrics for success
        ii. Results of the evaluations of scoring engines
    d. Implications for scoring approaches
        i. Summary of which engines can be used for which types of items
        ii. Descriptions of "gaps" in the scoring technology and mechanisms for filling those gaps by using human scoring
    e. Detailed recommendations for comprehensive scoring approaches
    f. Promising paths for continuous improvement over time.

Optional Establishment of an Open-Source Organization to Support Sustainability

The history of open-source software holds important lessons on sustainability. Richard Stallman is universally credited with starting the "free software" movement. As a programmer in MIT's Artificial Intelligence laboratory starting in the 1970s, Stallman contributed to a number of important AI systems. In 1985 he founded the Free Software Foundation, which launched the GNU project (GNU is a recursive acronym standing for "GNU's Not Unix") through which he introduced the concept of "copyleft," an antithesis of copyright [8, 9, 10]. The GNU project, thanks largely to funding from MIT, produced many important results. To this day, the GNU General Public License is the most widely used free software license.

Simply making source code available and distributing it free with few real restrictions on use are not sufficient to sustain software. The early history of the open-source Linux software provides helpful lessons. Without a supporting organization and either funding or a robust software developer community, open-source software often founders. The development of a robust software developer community often requires a strong leader.

The original Linux kernel was written by Linus Torvalds beginning in 1991. It was based on the Unix operating system, which was developed by AT&T's Bell Labs in 1969. The Linux project relied on the rewrite of Unix from assembly language to C by Dennis Ritchie in 1973 at Bell Labs, as well as various compilers and text editors that had been developed by the GNU project at MIT.

Many versions of Linux exist, but the most commonly known versions run on desktops and Web servers. Funding for these systems often rely on commercial vendors. SuSE Linux is now supported by Novell, which supports the community version and sells a commercial version. Red Hat's business model revolves around providing expert support. Many other versions of Linux run on closed government systems and university research systems or are embedded in closed commercial hardware.

The Debian distribution, in contrast, is an entirely volunteer project based on three foundational documents: the Debian Social Contract, the Debian Free Software Guidelines, and the Debian Constitution. The Debian project arose in 1993 and was developed by Ian Murdock from the Softlanding Linux System, the first distribution compiled from various software packages. Softlanding Linux was weighed down by poor maintenance and bugs. It was not until Bruce Perens succeeded Murdock as the project leader in 1996 that the project took off. Perens expanded the programmer base from 40 to 200, broke up the base system (which had been maintained solely by Murdock) by handing off maintenance to multiple programmers, and wrote the founding documents.

Under this contract, AIR will develop an open-source multiple scoring engine, based on existing open-source tools. At the end of this contract, support and maintenance of those tools will fall to SBAC. One way to reduce the risk in this transfer of responsibility is to build an organization that will carry the institutional knowledge of the software forward past the end of this contract. With the right funding, support, and staffing, the open-source organization will be in a position to manage the community of developers that will be required to sustain the open-source tools being built.

As an option, AIR proposes to do the following:

- Establish a separate independent not-for-profit organization to manage the open-source tools that arise from this project
- Hire a small staff for that organization, including a director, a lead software developer, a second software developer, and an assistant. If this option is selected, AIR will engage this staff in the conduct of this project to ensure that it carries institutional knowledge forward.

As a not-for-profit organization, the new organization will have a mission and a board of directors, both of which we suggest derive from SBAC. The mission of the organization will be to sustain and expand the open-source assessment products that arise from this contract and potentially from all SBAC activities.

Without this option, the software will be open source, and AIR will deliver it to SBAC with documentation. The contract contains no provision for support beyond the contract terms.

References

Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. Cognitive Science, 2, 155-192.

Brown, J. S., Burton, R. R., & deKleer, J. (1982). Pedagogical, natural language and knowledge engineering techniques in SOPHIE I, II and III. In D. Sleeman & J. S. Brown (Eds.), Intelligent Tutoring Systems (pp. 227-282). New York: Academic Press.

Burton, R. R. (1982). Diagnosing bugs in a simple procedural skill. In D. Sleeman & J. S. Brown (Eds.), Intelligent Tutoring Systems (pp. 157-183). New York: Academic Press.

Dobša, J. (2007). Comparison of information retrieval techniques: Latent semantic indexing (LSI) and concept indexing (CI). Varaždin, Croatia: University of Zagreb.

Chomsky, N. (1965). Aspects of the theory of syntax. Cambridge, MA: MIT Press.

Schank, R. C. (1975). Conceptual information processing. Amsterdam, The Netherlands: Elsevier Science Publishers.

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. Journal of Information Technology Education, 2, 319–330.

Dikli, S. (2006). An overview of automated scoring of essays. The Journal of Technology, Learning, and Assessment, 5(1), 1–36.

Hopcroft, J., & Ullman, J. (1979). Introduction to automata theory, languages, and computation. Reading, MA: Addison-Wesley.

Martin, W. A., & Fateman, R. J. (1971). The MACSYMA system. Proceedings of the second ACM symposium on symbolic and algebraic manipulation (pp. 59–75). New York: Association for Computing Machinery.

Stallman, R. (2011/1983). Initial announcement [GNU operating system]. Boston: Free Software Foundation. Available at http://www.gnu.org/gnu/initial-announcement.html

Free Software Foundation. (2011). What is free software? Boston: Author. Available at http://www.gnu.org/philosophy/free-sw.html

Free Software Foundation. (2011). What is copyleft? Boston: Author. Available at http://www.gnu.org/copyleft/copyleft.html

Part 3 Cognitive Labs and Small Scale Trials

L1 – 13 & 14 Plan and Conduct Cognitive Labs of New Stimulus Types, Item Types and Performance Tasks; Conduct Small-Scale Trials of Item and Task Types

Conduct research to try out new/innovative stimulus materials and item/task types across content areas, and grade levels to inform future item/task/stimulus development.

Introduction

Task 3 has two components: a series of cognitive lab studies and a series of small-scale tryouts, which will administer approximately 3,200 items to 50 to 200 students each. These activities must occur in time to inform large-scale item development. The success of these activities faces two substantial challenges:

- A pilot test during the 2012-2013 school year compresses the schedule for these research activities significantly.
- The delivery technology to be used in the pilot test will not be available in time for the small-scale tryouts.

The CTB/AIR/DRC/College Board team offers a solution that both meets these challenges and reduces other risks associated with the timely deployment of the pilot test.

The key features of our proposed approach follow:

- Solve the technology problem by proving AIR-developed item renderers under a broad open-source license without restrictions on use. Items are represented as XML documents with associated art and media. The items are rendered on the client-side testing system using our renderer. The renderer handles item display, interactions with the examinees, and capture and communication of student responses.
- Solve the timing problem through a four-part strategy based on SBAC objectives and RFP requirements. We begin by recognizing four categories of research questions that this task must address: (1) item content and structure questions; (2) stimulus format (type) questions; (3) item and stimulus layout questions; and (4) scoring questions. Among these, 1, 2, and 4 may provide feedback that affects item or stimulus content, whereas 3 affects only display. Therefore, answers to 1, 2, and 4 must be available before large-scale item development. Our strategy will allow us to do the following:

- Address the bulk of the item content and structure and stimulus format questions through cognitive labs, which can be completed by early summer, in time to inform most item development
- Recognizing that limits of existing (or soon to exist) scoring technologies restrict the types of questions that can be automatically scored (and therefore may influence SBAC item specs), conduct initial broad studies of the limits of automatic scoring using existing responses to constructed-response items to be contributed by SBAC states prior to large-scale item development
- Use the small-scale tryouts to confirm findings from the cognitive labs, systematically address presentation issues, systematically evaluate the effects of mixing item types and stimulus formats on tests, and gather responses with which to refine and evaluate automated scoring approaches
- Phase item development to develop items, stimuli, and pilot test forms to coordinate with the availability of the research results.

Under this plan, detailed below, the cognitive labs will be conducted starting immediately upon contract award, and run through the early summer. Item development for both the small-scale trials and the pilot test will take place over the summer, with the small-scale trials conducted in the early fall of 2012. Findings from the small-scale trials will inform the final phases of item development, which have been strategically designed to leave aspects of the stimuli and items that rely on this research until the fall. For example, the types of stimuli used or their particular structure, like other art, are typically described by item writers and created by artists later. This type of work will be delayed until after the small-scale trials whenever the trials are expected to inform their construction.

The open-sourcing of AIR item renderers will enable developers of the SBAC item authoring tool to integrate these renderers directly. Where the renderers diverge from SBAC preferences or specifications, developers of the authoring tool or the SBAC test delivery system may modify them to meet SBAC specifications. These renderers are all HTML-5 compatible and highly accessible.

This open-source strategy helps reduce timeline risk because these renderers exist and are mostly currently in operational use (our simulation renderer will be used in operational statewide tests this spring). The use of these renderers will ensure interoperability and will encourage competition (because they will be open-source and unrestricted). They have already proved to minimize bandwidth requirements, be easily integrated with multiple systems, and reduce the costs of developing technology-enhanced items. We integrate the same renderers (basically, stand-along Javascript modules with a well-defined API) with our test delivery system, our test development system, and our rangefinding system. SBAC contractors should have no problem with similar integration.

Using this approach, AIR can use our existing but proprietary test engine to deliver the small-scale tryouts (which will necessarily involve tens of thousands of students). This system has been proven in statewide high-stakes operation. By open-sourcing the item renderers, we eliminate any competitive advantage that this use of our system might bestow upon AIR. Incidentally, we believe that our renderers enable test developers to develop more cognitively challenging problem-solving types of items. Enabling all testing organizations to develop and deliver these types of items supports our not-for-profit mission of using the best science to improve people's lives—in this case through the improvement of education and assessment.

General Approach

Research Goals

The cognitive labs and small-scale tryouts have different objectives. In general, the cognitive labs are well suited to detect big effects or irregularities in content or processes that are new and relatively unstudied. Cognitive labs' use of small samples and in-depth questioning can help identify "overlooked" implications of new item types or stimuli. The small samples also make the studies susceptible to idiosyncratic response characteristics of particular individuals included in the sample. Therefore, the appropriate use of cognitive labs focuses on that which is common across many of the subjects.

The small-scale tryouts will involve a more representative sample, and a larger sample. The larger sample makes them sensitive to less universal reactions; however, they provide less opportunity to explore effects that have not been previously foreseen or hypothesized.

Research Goals of the Cognitive Lab Studies

The cognitive labs will focus on the new innovative item and stimulus types and on usability considerations for students who may have special access requirements. They will offer an opportunity to explore how these item and stimulus types function and to investigate whether and when specific innovations enhance measurement.

For the cognitive labs, we will write items that adhere to the item specifications (to be published in February) and to the guidelines set forth there. In the absence of SBAC technology, we plan to deploy these items in AIR item renderers as they exist. Therefore, there may be some aspects of the SBAC specifications that are not mimicked exactly, but we will endeavor to match the specifications as closely as possible.

Where the specifications leave room for innovation, or where SBAC expresses concerns about the specifications, we will draft items that innovate in known ways so that student interactions with these items may be compared with interactions with their less innovative counterparts.

The cognitive labs will focus on

- items and stimuli designed to measure, or contribute to the measurement of, previously hard-to-measure constructs, especially those that at the greater depth of knowledge;
- accessibility of different stimulus formats and types and response mechanisms (e.g., drag and drop, hotspot);
- technology-enhanced stimuli; and
- the impact of accessibility features and options.

We anticipate that much of the innovation in the item specifications will be for harder-to-measure constructs, including both items and performance tasks. For each example item or task studied, we will identify the learning target and the nature of the evidence that the item is intended to provide about that learning target. By monitoring the students' think-aloud responses and by using directed interviews with the students following their responses to the questions, we will ascertain whether students generally employed the anticipated cognitive processes. We will look across subjects and items intended to activate similar processes for patterns in the types of cognitive processes that are successfully or unsuccessfully activated. Similarly, we will look for patterns among the successful and unsuccessful measurement strategies employed by the innovative items and stimuli.

Through traditional usability analysis, the cognitive labs can contribute meaningfully to the accessibility and usability of the various stimulus formats and types of response mechanisms. In this regard, even standard multiple-choice items are of interest. The layout on the page, the arrangement of response options, visual cues when mousing over options, and other features will affect students' ability to respond. Increasing the usability of the items reduces the construct-irrelevant variation in students' responses. Although this portion of the study will include standard multiple-choice items, it will focus on technology-enhanced response mechanisms and mechanisms for collecting constructed responses.

We anticipate an increased reliance on technology-enhanced stimuli. In some cases these may be animations or audio. In other cases they may be interactive stimuli that the examinee may manipulate. The cognitive labs will explore the following questions, among others:

- Do students use these stimuli as anticipated (e.g., do they watch the animations? If so, how many times? Do they replay the animations before answer some or all questions?)
- Are students able to access the relevant information in the stimuli, or are the format or other characteristics distracting?
- Can students answer the questions without referring to the stimuli or parts of the stimuli?

Where relevant, we will explore where the addition of technology-enhanced stimuli may increase access for some students (e.g., audio may reduce barriers for limited-English students) or possibly raise barriers for others (e.g., graphic-intensive stimuli may impede students with visual impairments). Where technology-enhanced stimuli seem to improve measurement validity or reliability for the majority of students but raise barriers for others, we will investigate the mitigating effects of accessibility tools and accommodations.

Part of the usability of the system reflects how students use the options available to them on the system. Some research suggests that providing configuration options poses unnecessary barriers and frustrations to users (Ceaparu et al., 2004; Schniderman, 2004), while others believe that it increases access. The cognitive labs will evaluate whether usability considerations favor offering configuration options during the test or whether these accessibility settings are better established before the student begins the test.

This portion of the investigation will investigate (1) whether students access accessibility features when they might be helped by them, (2) whether the features appear to alter the cognitive processes employed, and (3) whether students are distracted by the accessibility tools.

Research goals of the Small-Scale Tryouts

The objectives for the small-scale trials described in the RFP can be categorized under three general themes:

- To gather the student responses needed to inform scoring, both automated and human
- To "try out" the tasks to get early indications of whether findings from the cognitive labs hold true with larger more diverse samples and whether students generate the range of responses expected
- To systematically evaluate presentation questions, including comparing different layouts and formats for items and stimuli and evaluating the impact of different mixes of item and stimulus types on tests.

As mentioned in the introduction to Part 2, information about automated scoring may provide relevant feedback to the test specifications and item development process. For example, some scoring engines function quite well with a constrained domain, and SBAC may want to consider such constraints for some subset of item specifications intended for automated scoring. This sort of information will be valuable early in the development process, and AIR proposes to conduct the initial research that will yield these sorts of general insights, using existing student responses donated by SBAC states for this purpose.

During the actual small-scale trials, we will, of course, gather the information needed to fine-tune the scoring engines and the scoring rules for both human and automated scoring.

The small-scale trials will provide a more systematic test of some of the findings from the cognitive labs, and they will help us investigate whether the items, and particularly the innovative items, function consistently across identifiable groups. Recall that our strategy is to move forward with item development on the basis of results from the cognitive labs. There is some risk that findings from the small-scale tryouts may identify some differential item functioning across groups or fail to uphold the cognitive lab findings.

Our plan accommodates potential findings contrary to those from the cognitive labs. In these cases, items may require modification during the review process—an opportunity that we will preserve. Readers should also recall that the development of items and stimuli that depend most on findings from the small-scale tryouts will be delayed until those results are available. This plan strikes the balance necessary to exploit the potential of these research opportunities while maintaining the SBAC pilot test schedule.

The final set of questions about the presentation of items and stimuli has at least four aspects:

- The particular media (format or type) used to present stimuli
- The layout of the presentation
- The effect of mixing item and stimulus formats on a single test
- The interaction of the stimulus format and available accessibility tools.

To address these questions we propose to develop different variants of items and stimuli, varying layouts and media in a true experimental design. This process will allow us to detect factors that influence student performance and, therefore, the validity of the items.

Here are a few illustrative examples:

- The student is presented with a stimulus in the form of a "how to…" informational passage about building a bird house and a set of items. The questions range from simple comprehension to bigger picture ideas. Another version of the same stimulus includes an accompanying video or animation that shows the steps being followed to actually build the bird house.
- The layout of a selected-response item is presented in a format that embeds the graphic in the stem and lists the options vertically underneath the question. The same item can also be presented with the graphic to the left of the stem and the options listed vertically to the right of the stem (many different layouts are available for selected-response items with and without graphics).
- The layout of a reading stimulus is presented in a split-screen format with the passage on the left of the screen and the items on the right. In another variation, the passage appears in a pop-up window that can be moved around and resized, with the items appearing in the stationary window beneath it.
- A selected-response item is presented in the basic multiple-choice format, and the same content is also presented using hot spot technology. For example, the student is asked to identify a detail from the passage that exemplifies personification. The item can be presented as a multiple-choice question offering four options with three distinct distractors or with hot spot technology where the student selects the details from within the actual passage.
- An interactive stimulus is presented for the student to enter different inputs into a function machine that generates outputs. The student is required to determine the function that is being used. This could also be assessed using a static table with a few values of inputs and outputs.
- A technology-enhanced item that requires the student to plot points, generate a line of best fit, and make a prediction for a new x-value based on the line of best fit is presented in two different ways. One way may present a bulleted list with very explicit directions for the student:
    - Use the Add Point tool to plot the given points in the data table.
    - Use the Add Line pool to generate a line of best fit.
    - Predict the y-value when the x-value is 5.
    - The second version may be left very unscaffolded. The student is still given all the same tools but the task would be worded differently:
    - Predict the y-value when the x-value is 5. Justify your prediction.

Timeline

The timeline depicted in Figure 11 schematically illustrates the phasing and integration of the research and item development for this project. Initial results from the cognitive labs will investigate the largest, riskiest innovations, and return results immediately for use in the development of pilot-test items and inform the study design for the small-scale trials.

Item development for the pilot test will defer aspects of development until key results are available. The small-scale trials confirm (or contradict) findings from the cognitive labs, and inform decisions about media used in stimuli, presentation of items in different formats, and the use of different layouts. These aspects are readily deferred where necessary until findings from the trials are available.

Figure 11: Research and Item Development Phasing

Item Development for Cognitive Labs and Small-Scale Tryouts

Introduction and Open-Source Strategy

The cognitive lab studies must begin almost immediately upon contract award. These studies will necessarily evaluate innovative item types, as well as the usability of the technology-enhanced items. Item development for the small-scale tryouts cannot lag far behind—these are scheduled for early fall 2012 and must try out nearly 4,000 items and tasks. If each item is to be administered to 150 to 200 students, these "small scale" tryouts may involve more than 30,000 students and therefore must rely on robust technology.

To meet these aggressive timelines, minimize risk, and maximize the opportunity for innovative measurement, we propose to use AIR's test delivery system to deliver items for the cognitive labs and the small-scale tryout. Our test delivery system has the following benefits:

- Is in operational use in four states, providing assurance that it can be successfully used to test large numbers of students in actual schools
- Offers a broad range of accessibility options, ensuring that SBAC will have the opportunity to evaluate the impact of these options in both cognitive lab settings and the small-scale tryouts
- Supports a robust array of innovative item types, including
    - Hot spot
    - Drag and drop
    - Drawing
    - Matching
    - Assembly
    - Graphing
    - Gridded response
    - Natural language constructed response
    - Essays
    - Simulations

- Supports virtually all Web-deliverable media, including, for example, audio, video, and animations
- Can support the deployment of innovative items almost immediately upon contract award.

Relying on an existing proprietary system for these research studies has the advantage of reducing risk. We recognize that this also risks two disadvantages: the capabilities of the system may not exactly match those for which the item specifications have been designed, and the proprietary nature of the system risks reducing competition.

Given the timelines, it is not possible to extend the capabilities of AIR's system between the delivery of the specifications at the end of February and the beginning of the item development contract one month later. However, our system is highly flexible, our innovative item technologies are robust, and our system supports many accessibility features. We have more than 4,000 innovative items in operational use in statewide tests, representing a wide variety of capabilities. We will be pleased to demonstrate these capabilities, and we are confident that the existing capabilities will meet most of the SBAC specifications and that this will provide the most complete solution available.

We offer an innovative solution to alleviate any concerns about competition. If awarded this contract, AIR will release our item renderers under an open-source license that allows unrestricted use. This will provide SBAC the ability to integrate these renderers into SBAC systems (including your item development system, rangefinding system, and test delivery system) without incurring additional cost. Other vendors will have the opportunity to integrate these renderers, innovate, and incorporate innovations into the software.

Alternatively, the items can be rendered to QTI and administered through any QTI-compliant renderers. We note that QTI does not offer robust native support for assembly, drawing, some types of graphing, or simulation formats. Where these item types are supported, it is through custom external interfaces and tools, which are not themselves part of the standard.

We draw particular attention to our simulation technology. The simulation technology puts powerful simulation tools in the hands of test developers. The simulators can capture and return key student interactions, display data in both tables and graphs, and be paired with items of all types. The system minimizes the role of animators and others and empowers the test developers—those with the expertise—to directly enter the calculations and logic that drive the simulations. Imagine, for example, a mathematics item that asks the student to enter values for various parameters and allows the student to see the data graphed. Associated questions might ask the student to write or identify equations that approximate the graphed function. (Incidentally, the simulation described can be implemented by the test developer alone, without any animator—the graphing functions are built in.)

These renderers, including the simulation renderer, would be released under an open-source license.

Item and Task Development for the Cognitive Labs

Item and task development for the cognitive labs will follow the SBAC item specifications due out at the end of February. As specified in the RFP, the cognitive labs will study 40 performance tasks and 480 items and stimuli. We propose that the distribution of items across grades be approximately proportional, with approximately two or three performance tasks per grade/subject and approximately 36 items per grade/subject.

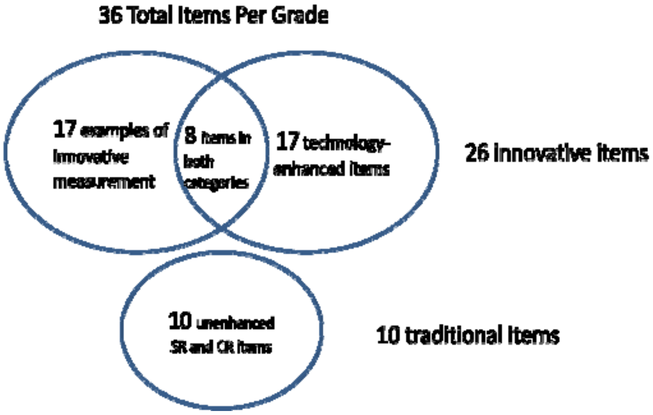
Recall from above that we identified four classes of research questions to be addressed by the cognitive labs:

- Whether innovative items and types elicit the intended cognitive processes
- The usability and accessibility of technology-enhanced and other items
- The accessibility of various stimulus types
- The impact of accessibility features and options on student access and validity.

In developing items for the cognitive labs, we anticipate that approximately half the items (and all performance tasks) will fall into the first category. Cognitive labs provide the most value when investigating novel approaches.

We also anticipate that approximately half the items created will include technology-enhanced stimuli or response mechanisms. That said, we expect that item specifications will dictate that technology-enhanced items will be more often associated with harder-to-measure constructs, so we anticipate significant overlap between these two categories. If we assume that approximately half the technology-enhanced items will target harder-to-measure constructs, then approximately 10 of the items created will fit into both categories. The distribution of items to be developed in each grade is illustrated in Figure 12.

Figure 12: Distribution of item types developed for cognitive labs


36 Total Items Per Grade
17 examples of Innovative measurement
8 items in both categories
17 technology-enhanced items
26 innovative items
10 unenhanced SR and CR items
10 traditional items

If admissible under the tasks specifications, we recommend that at least one task per grade be an extended online simulation to test the feasibility of this mode of assessment. Many of the example tasks presented in the draft specifications on January 6 are amenable to machine delivery and potentially to automated scoring for major sections.

In each grade and subject, at least one of the items and tasks will be taken directly from the examples in the specifications. There are several valid reasons for using the items and tasks from the specifications:

- The items were developed to represent exemplar item wordings and formats for assessing the content delineated by the standards.
- The items and tasks in the specifications have all been vetted and approved by the SBAC content committees.
- The specifications will be strengthened by the inclusion of these items in the cognitive labs and small-scale tryouts, because SBAC will have data and experimental evidence that confirms the usefulness of these items for the assessment of the constructs in the Common Core State Standards.

Item and Task Development for the Small-Scale Trials

We propose to design the small-scale trials to meet three main objectives:

- To confirm that the inferences draw in the cognitive lab studies hold up in larger more diverse samples and that innovative features of the items function fairly across identifiable groups of students
- To systematically investigate questions of presentation and technological enhancement

- To gather the information needed to test and validate the automated-scoring approaches for constructed-response items.

Accomplishing these three objectives will require careful engineering of the items developed for these studies.

Item banks for the small-scale trials should support:

- direct comparison of items and stimuli using varied presentations, response technologies, and stimulus formats;
- validation of inferences about what students know and can do based on responses to varied items; and
- validation and refinement of automated-scoring approaches.

To accomplish these objectives, the small-scale trials ought not to seek an item bank that will measure the full breadth of the Common Core State Standards but rather focus on those where innovative measurement and the use of technology offer the greatest opportunities for advancement.

Matched item sets

Within each grade and subject, we propose to develop matched sets of items. The items in each set will include

- conventional selected- and constructed-response items,
- technology-enhanced items and stimuli, and
- technology enhancements to the response mechanisms (e.g., hot spot, drag and drop).

These sets of items will allow a clear and systematic comparison about the response characteristics associated with each delivery mode or feature. Differences across members of the set will provide a basis for inferences about key questions, such as the conditions under which hot-spot items provide more (or less) valid measurement.

The inclusion of multiple modes of measurement for each construct provides an opportunity to validate inferences about what students know and can do. Students who demonstrate knowledge or skills using one mode but who fail to successfully demonstrate the same or closely related skills or knowledge under a different mode provide evidence against the validity of the latter mode. Where differences across modes exist, we look across different items to identify the situations that favor each mode for valid measurement.

We propose to develop sets of items that will enable us to answer a variety of questions, some of which exist now and others which may arise during the cognitive lab or process of test development or be raised by the specifications documents. Examples of the types of issues that item sets might be created to address follow:

- What factors, if any, should influence the choice of response mode for selected-response items? Many hot-spot, drag-and-drop, ordering, and other item types could be represented by multiple-choice questions. Under what conditions does each response mechanism minimize construct-irrelevant variance? For example, hot-spot items might increase usability when the response is inherently about visual elements that exist naturally in positions relative to other visual elements that serve as distractors.
- What mechanisms for presenting stimuli and directions reduce construct-irrelevant variance or otherwise have an impact on the validity of the intended inferences? For example, does a visual emphasis on key words in passages reduce barriers to comprehension or does it cue responses? What is the impact of including introductory statements for stimuli or using bulleted lists for directions? Item sets can be designed to vary these types of features to inform whether and when to use each.
- When is construct-irrelevant variance reduced through the use of audio (rather than text), video (rather than still graphics), and interactive stimuli (rather than stimuli that are passively viewed)?

Accessibility

We recognize that each innovation brings with it risks or benefits for students with different capabilities. For example, an increased use of visual stimuli (animations, video, etc.) and graphic response mechanisms may reduce barriers for students with limited English proficiency or with disabilities affecting their reading skills, but increase barriers for students with visual impairments. The small-scale trials will provide an opportunity to study the effectiveness of access tools and accommodations at reducing or eliminating these disparities. Specifically, by measuring the same content in different ways, we will be able to identify discrepancies in performance across response modes that persist with the available access tools and accommodations.

Constructed-response items, scoring rules, and automated scoring

The constructed-response items in the item sets will perform double duty. First, as members of the item set, they will contribute to the study of item validity, enhancements, and accessibility for the rest of the items in their sets (as well as for themselves). Second, taken individually and across all constructed-response items, they will contribute to the development of scoring rules, investigations of scoring approaches, and the refinement and validation of automated-scoring approaches.

All constructed-response items in the study will be double-scored, and a resolution score will be assigned by an expert scorer when there are discrepancies. Approximately one third of the constructed-response items will repeat this process using a different scoring approach (e.g., checklists rather than traditional rule-based training).

The constructed-response items will also be scored automatically using the approach deemed most appropriate for each item. Recall that a significant amount of research on the scoring engines will be conducted prior to the small-scale trials using existing items and responses contributed by member states. The small-scale trials will evaluate whether the findings from that research hold when used with items designed to measure the Common Core State Standards using the SBAC item specifications.

The RFP indicates that research suggests that 150 to 200 responses per item should be sufficient to refine and validate the scoring engines. This has not been our experience and is not typical for black-box scoring engines. Black-box engines represent most of the AI scoring engines in use for essay scoring. In general, at least 150 to 200 responses are necessary to train the engine. Often a larger sample is necessary to obtain a sufficiently representative sample at all score points. Training such engines is only the first step. Best practice requires an independent sample for use in the validation of scores. Black-box engines tend to be highly dimensional, with many parameters that depend on the training sample. In almost all cases, this results in better fit to the training sample than is found when extending beyond the training sample. This requires that validation take place on a set of responses not used in the training.

We can price an option that goes beyond the scope defined in the RFP by significantly expanding the sample of responses to constructed-response items that will be used to evaluate the constructed-response scoring engines.

Performance tasks and scoring models for performance tasks

As part of the Vision Document to be created (see Part 2), we will outline various options for scoring performance tasks. Under this contract, we intend to evaluate the efficacy of automated-scoring models for the tasks and to develop options for how automated scoring or hybrid scoring can be used with the tasks. The sample tasks released with the January 6 draft of the specifications hold promise for automated-scoring and hybrid-scoring models.

CTB, AIR, and our other partners will develop approximately 64 performance tasks across grades and subjects. These tasks will cover a range of types of tasks, varying the extent to which resources are made available to students (versus student-identified) and tools are prescribed or left to student choice, as well as administration mode, level of scaffolding, and other factors thought to influence the skills elicited by the tasks.

The specific research questions to be investigated cannot be solidified until the item specifications are available. However, we present an illustrative item development plan designed to support the investigation of some likely research questions.

We propose to develop items in sets of six items measuring the same content using different modes. Three items, two multiple-choice and one-constructed response, will measure each construct using traditional methods. The other three items in the set will represent technology-enhanced items, each of which enhances the single item in different ways (see Figure 13).

Figure 13: Illustrative Composition of Item Sets

This design will yield an item bank and set of forms that are sufficient to address the range of research questions required. Specifically, the 36 sets will ensure the opportunity to evaluate many different configurations. For example, 12 of the sets might compare three different response mechanisms, 12 of the sets might compare three different stimulus types/formats, and the remaining 12 might evaluate different directions, layout, and other presentation options. Across two subjects, this design provides opportunities to evaluate at least 18 variations per grade.

In fact, 12 sets testing the same concepts are unnecessary. If each item in each set is seen by 200 students, this will yield 2,400 responses that can be aggregated to address each comparison. SBAC may, for example, want to use six item sets per grade subject to compare hot-spot, drag-and-drop, and multiple-choice response mechanisms, and another six item sets to compare drag-and-drop, ordering, and matching items.

Under this design, we recommend targeting about 10 to 20 learning targets per grade/subject, focusing on those that require the deepest cognitive skills and those that have been the most difficult to adequately measure in the past.

We strongly encourage SBAC to use the small-scale trials to investigate the conditions under which each of the various selected-response mechanisms is appropriate. Currently, little research exists on the relative merits of drag-and-drop, hot-spot, matching, ordering, and multiple-choice items. All represent variants of selected-response items, and most examples found on state tests represent a selection among relatively few choices. As SBAC moves toward a more diverse set of item types that may all be used to represent the same content, it will need a research base to help identify the factors that recommend the use of one representation over another. Therefore, we may want to vary the response mode for similar items measuring the same content. As mentioned above, this design supports similar evaluations of different stimulus types and presentation methods.

Another critical research question is whether mixing response types affects the validity of inferences from student responses. In constructing forms from the item sets, we can construct forms that are either homogeneous or heterogeneous with respect to item types, stimulus formats and types, and other presentation characteristics. (Recall that each item set will be split across four forms.)

The final set of research objectives of the small-scale trials is to inform the scoring of constructed-response items. This includes evaluating varied methods of hand-scoring, training and refining automated scoring engines, and testing different scoring models (e.g., models in which scores that are particularly important to final classification or are automatically scored with high uncertainty are routed to human scorers for verification).

This design will yield approximately 35 to 40 constructed-response items per grade per subject. If constructed as suggested in Figure 13, the constructed-response items will each receive twice as many responses as the other items, or a total 300 to 400 responses per item. This begins to address the possible deficit in responses, because most black-box scoring approaches require at least 150 to 200 responses to train (and often more), and best practice requires an independent sample for validation of the scoring.

Performance task plan

As with the item development plan, we cannot specify the specific research questions to be investigated until the performance task specifications are available. However, we present an illustrative task development plan designed to support the investigation of some likely research questions based on the criteria for performance tasks outlined in the RFP.

Our goal will be to use the SBAC performance task specifications to create meaningful complex performance tasks that align with multiple strands or domains. The tasks will enable us to assess higher-level thinking skills and higher depths of knowledge that are difficult to assess in more traditional multiple-choice and constructed-response testing formats. Tasks may allow students to interact with multiple stimuli and require multistep planning to manage the analysis and synthesis of ideas. Although the task development process will parallel the item development process, the complex nature of performance tasks demands that we emphasize the importance of the following qualities:

- Alignment—Each part of the task aligns with the content standards and task specifications.
- Meaningfulness—The task allows the performance of real-world tasks and 21st century skills.
- Accessibility—All directions and stimuli are clear and accessible.

Practicality—The information obtained from the student is worth the time spent on the task.

The cognitive labs will help evaluate how the tasks function, and we will use the results to inform the small-scale tryouts. The goal of the tryouts will be to further refine our approach to determine whether students generate the range of expected responses.

By nature, performance tasks are more complex than even the highest complexity items. Performance tasks integrate knowledge across standards and demand complex analysis and higher-level thinking. Because performance tasks often ask students to analyze and synthesize information, they also often demand that students produce more extended responses that lead to more complex scoring criteria. Thus, while some performance tasks such as writing prompts can be straightforward, they often involve "more": more and longer stimuli, more complex directions, and more complex scoring rules. The complex nature of performance tasks lends itself to multiple approaches to task formatting, presentation, and scoring.

As with the item development, we propose to develop sets of tasks that will enable us to answer key questions, some of which we can posit now, others that will be raised by the specifications documents, and others that arise during the cognitive lab and test development process. Examples of the types of issues that performance tasks sets might be created to address follow:

- What factors enable students to focus on the standards-based knowledge being assessed, and minimize construct-irrelevant barriers?
- When do the tasks adequately result in students' accessing the cognitive processes and depth of knowledge we expect?

- Which methods of presenting directions and stimuli make it more likely that students access and understand the entire task?
- Which configurations of technology-enhanced items allow robust student engagement and response?
- What types of scoring criteria allow tasks that are fairly and consistently scored?

As specified in the RFP, the small-scale tryouts will test 64 performance tasks. We propose to develop four or five performance tasks per grade per subject and to test each of the 64 performance tasks with 150 students. The performance tasks will support the objectives listed in Section 3.3.3.

As with the plan for item development, we propose to develop sets of performance tasks measuring the same content. Tasks measuring similar content and concepts will feature different presentations, different response mechanisms, and different scoring. We propose making such comparisons both within and across grade levels. At least one task per grade will be completely online—all the stimuli and information needed by the student will be available online, and the item will be scored online.

One critical research objective might be to investigate the presentation of the performance task. We will evaluate varied methods of presenting stimuli, directions, and the amount of scaffolding provided to the student. Stimuli may be presented as text, video, audio, or a combination. Directions may be provided at the start of the task or throughout the task. A task may be presented all at once or "chunked," with the student performing one part of the task, then moving to another part, then ending with a synthesis of the whole.

As with the plan for item development, we encourage SBAC to investigate the conditions under which each of the various response mechanisms is appropriate. We propose to incorporate drag-and-drop, hot-spot, and simulation mechanisms into performance tasks. We may create similar performance tasks but vary or mix the response mechanisms within each of those tasks. In evaluating technology enhancements to items, we may also vary the way students interact with different stimulus types.

Another critical research objective is to inform the scoring of performance tasks. We recognize that different performance tasks will require different scoring models, and hope to investigate several of these models. For each grade, we will create at least one task that uses automated scoring. Other tasks may be independently scored by teachers, with a percentage audited to evaluate accuracy. We may also use electronic scoring, routing tasks to teachers in different states and asking them to use scoring guides to enter scores electronically.

We may also investigate the effect of more than one of these issues in a single task when appropriate. Consider a highly scaffolded performance task, in which a student is asked to complete several smaller tasks before synthesizing them into a main task. The primary point of study would be the effect of this scaffolding on the student's ability to create an effective response to the main task. However, we may decide to score each of the smaller tasks, incorporating the scoring criteria for these tasks into the scoring criteria for the task overall. The additional scoring criteria may provide us with richer information about what the student knows and is able to do at that content standard.

In addition to comparing tasks within grade, we propose to compare tasks across grades. When possible, we may create similar tasks to investigate the effect of such things as task presentation and response type across grades. In rare instances, we may measure similar skills across grades: consider an expository writing prompt for which we vary the directions, amount of scaffolding, and presentation of the item across multiple grade levels. The task and scoring criteria would be very similar across grade levels. However, because many standards are not consistent or comparable across grades, it will be difficult to create tasks of sufficient complexity that work at multiple grade levels. We propose creating design patterns or task templates that carry different presentation structures, response types, or scoring models across grade levels. These patterns/templates will be based on the task specifications and the information yielded from the cognitive labs and the task development process.

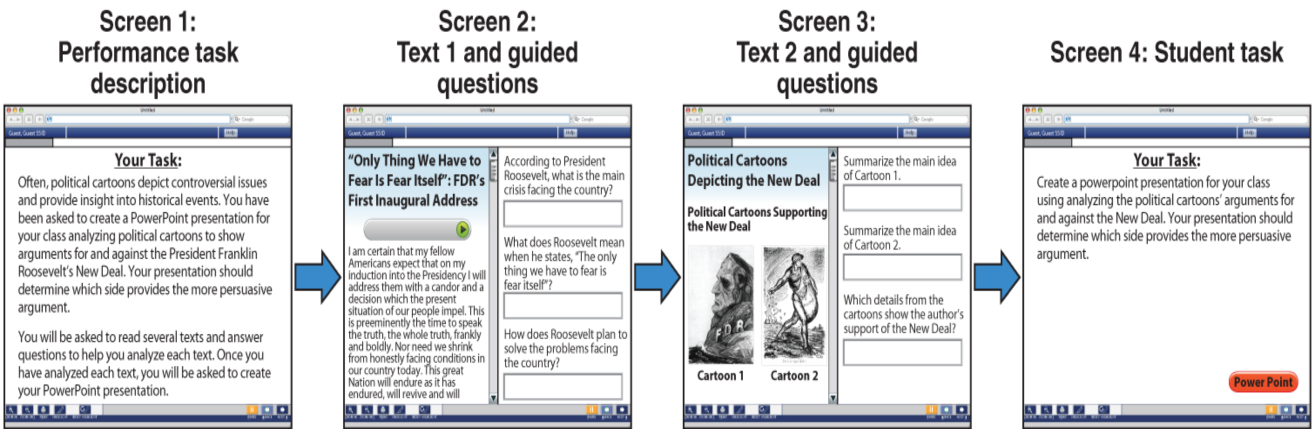
The design patterns/task templates will provide strategies for consistent approaches to presentation, response type, and scoring issues when developers create performance tasks. For example, we might create strategies for when and how to scaffold tasks and directions. This design will provide information about the grade level at which students respond effectively to a mix of response types, the types of scaffolding that might be effective at different grade levels, or the grades at which scoring scaffolded tasks provides valuable information. See Figure 14.

Figure 14: Performance Tasks Across Grade Levels

	Grade 3	Grade 4	Grade 5	Grade 6
Task 1	Focus on main task Score main task only	→	More open-ended Score main task only	
Task 2 (variation of Task 1)	Highly scaffolded tasks and directions Score each part as well as main task		Scaffolding for more complex tasks with fewer directions Score each part	
Task 3	One response type Use automated scoring	→	More sophisticated use of technology-enhanced items Use automated scoring	
Task 4 (variation of Task 3)	Mix of response types Use human scoring		Mix of more sophisticated response types Use human scoring	

The two variations of item presentation in Figures 15 and 16 show a scaffolded approach to the performance task. In the first version, the task is broken up into four screens. Each screen represents a different part of the task. Screen 1 outlines the task for students. They then click to screen 2, where they find the first text for analysis. Students read the text and answer guided questions that help them think through the issues. Students then move to screen 3, where they read the second text and answer guided questions to help them analyze the issues raised. Once students have worked through these scaffolded tasks, they move to screen 4 to complete the overall performance task.

Figure 15: Performance Task Tryouts: Varying Presentation



Screen 1: Performance task description

Your Task:

Often, political cartoons depict controversial issues and provide insight into historical events. You have been asked to create a PowerPoint presentation for your class analyzing political cartoons to show arguments for and against the President Franklin Roosevelt's New Deal. Your presentation should determine which side provides the more persuasive argument.

You will be asked to read several texts and answer questions to help you analyze each text. Once you have analyzed each text, you will be asked to create your PowerPoint presentation.

Screen 2: Text 1 and guided questions

"Only Thing We Have to Fear Is Fear Itself": FDR's First Inaugural Address

I am certain that my fellow Americans expect that on my induction into the Presidency I will address them with a candor and a decision which the present situation of our people impel. This is preeminently the time to speak the truth, the whole truth, frankly and boldly. Nor need we shrink from honestly facing conditions in our country today. This great Nation will endure as it has endured, will revive and will

According to President Roosevelt, what is the main crisis facing the country?

What does Roosevelt mean when he states, "The only thing we have to fear is fear itself"?

How does Roosevelt plan to solve the problems facing the country?

Screen 3: Text 2 and guided questions

Political Cartoons Depicting the New Deal

Political Cartoons Supporting the New Deal

Cartoon 1 Cartoon 2

Summarize the main idea of Cartoon 1.

Summarize the main idea of Cartoon 2.

Which details from the cartoons show the author's support of the New Deal?

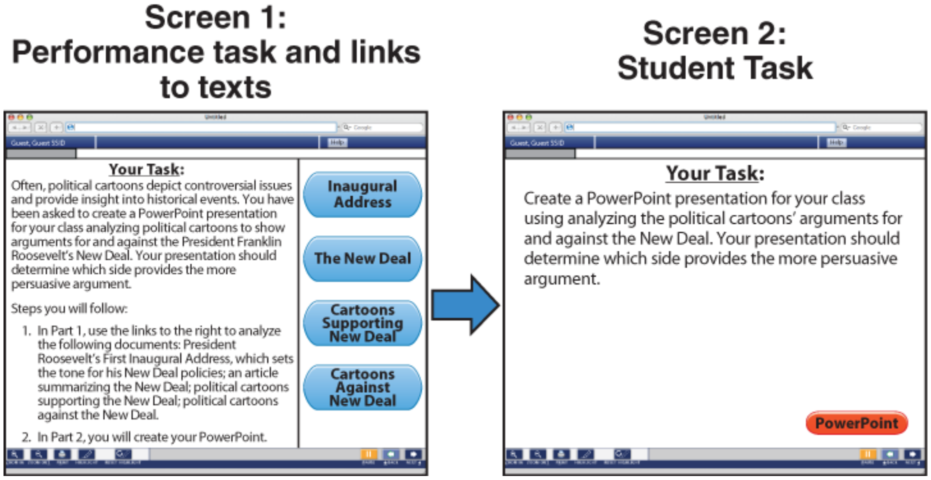
Screen 4: Student task

Your Task:

Create a powerpoint presentation for your class using analyzing the political cartoons' arguments for and against the New Deal. Your presentation should determine which side provides the more persuasive argument.

The second variation of item presentation asks students to complete the tasks in a more independent unstructured manner. The first screen describes the task and provides a link to the texts that students will analyze to complete the task. Clicking each link will allow students to access each text. The second screen restates the task and allows access to PowerPoint slides necessary to complete the task.

Figure 16:



Screen 1: Performance task and links to texts

Your Task:

Often, political cartoons depict controversial issues and provide insight into historical events. You have been asked to create a PowerPoint presentation for your class analyzing political cartoons to show arguments for and against the President Franklin Roosevelt's New Deal. Your presentation should determine which side provides the more persuasive argument.

Steps you will follow:

1. In Part 1, use the links to the right to analyze the following documents: President Roosevelt's First Inaugural Address, which sets the tone for his New Deal policies; an article summarizing the New Deal; political cartoons supporting the New Deal; political cartoons against the New Deal.
2. In Part 2, you will create your PowerPoint.

Screen 2: Student Task

Your Task:

Create a PowerPoint presentation for your class using analyzing the political cartoons' arguments for and against the New Deal. Your presentation should determine which side provides the more persuasive argument.

Item Development Process for Cognitive Labs and Small-Scale Trials

Introduction

Our team will develop items for the cognitive labs and small-scale tryouts using a thorough internal and external review process. The item development will proceed on a rolling basis. The outcomes from focus groups, cognitive labs, and small-scale tryouts will drive item development. Each of these research events will yield information about the items, which in turn will contribute to item validity. AIR proposes to use focus groups during the initial phase of development to obtain educator input on the proposed innovative item formats. We envision the results of these focus groups as helping to shape the issues that should be addressed in the cognitive labs.

Our team will use CTB's item development tool CIAS as a database that maintains each version of the item with all associated stimulus and item specific attributes. In some cases, where item types are needed that are not supported by CIAS, we may supplement this tool with others. CIAS will also facilitate the linear review process that AIR will use when developing items. The reviews conducted on each item will include a series of AIR internal reviews as well as SBAC member reviews, including Bias/Sensitivity.

Focus Groups

AIR proposes to conduct focus groups shortly following the award of the contract. The focus groups will include representative educators from the SBAC states. The purpose of the focus groups is to obtain feedback from teachers about the different types of items and innovations that will be used on the SBAC assessments. The data will inform the development of the items for the cognitive labs and help structure the student observations conducted during the cognitive labs.

The participants for the focus groups will be recruited on the basis of SBAC participation guidelines and qualification criteria in the involvement of individuals to serve on each review committee. The focus group will review sample items and stimulus presented in different formats using multiple innovative types. The participants will be asked to comment on the items and stimulus and the processes they think will be used by students to complete the items. Focus group members will give feedback about the way the content is currently presented. They will also have a chance to offer varying presentations and formats that may work well with certain subjects, constructs, or grade levels. A second outcome of the focus group will be to obtain ideas about how students are likely to interact with item and stimulus types.

Because the cognitive labs will focus on items and stimuli meant to measure previously hard-to-measure constructs, the focus group will also look at items and stimuli measuring these constructs. The members of the focus groups will discuss the benefits of using different stimulus formats, types of response mechanisms, and technology enhancements for these constructs. Accessibility features and options will also be discussed. Examples such as the following will be presented to the group:

- Items with varying placement of options
- Items with varying visual cues embedded in stem/stimulus
- Items with varying response mechanisms including but not limited to drag and drop, hot spot, drawing, matching, and graphing
- Stimulus presented as static information, table, graph, passage
- Stimulus presented in different positions relative to the items (to the left, on top, pop up).
- Stimulus presented as different media (audio/video)
- Items and stimulus with text to speech options and color overlays

One purpose of the cognitive lab is to determine the mental process that students are using to complete an item. We will compare the anticipated use or cognitive process with the processes that students actually use. Discussions with the focus group will help us determine what the anticipated cognitive processes are for different constructs and items or stimulus types. For example, the focus group may assert that students will extend a given line graph before answering a question about a prediction. During the cognitive labs, this type of assertion can be assessed by asking students to explain their process for answering the question. In addition, it may be possible to offer an innovative way for students to use technology to extend the graph, which could also be tested in the cognitive lab environment. Information gained from the focus groups will be invaluable to the cognitive lab process.

The item development for the cognitive labs and ultimately for the small-scale tryouts will be shaped by the outcome of the focus groups. Having input from educators at the beginning of the development process will yield valuable information for all phases of the development process.

Item Development Process

The item development process consists of writers, content reviewers, editors, bias/sensitivity and accessibility reviewers, and SBAC reviewers. Rigorous internal reviews of all items, item classifications, and scoring rubrics will be employed. Throughout the review process, a record of all items' relevant attributes and reviews will be maintained in the item development system.

Item development begins with clear guidelines that include Item/Task Specifications, Stimulus Specifications, Accessibility and Accommodations Guidelines, Bias/Sensitively Guidelines, and editorial style guides developed by the contractor from RFP 04. The principles of universal design will also be applied during item development. These guidelines ensure that each aspect of an SBAC item is relevant to the measured construct. They help ensure that the wording, required background knowledge, and other aspects of the item are familiar across identifiable groups. These guidelines will pervade our item development process from item inception through internal and external reviews.

We propose that all newly developed items for use in the cognitive labs and small-scale tryouts go through the following internal reviews and external reviews:

- Internal Reviews: Content, Editorial, and Senior Content sign-off
- External Reviews: Bias/Sensitivity and/or SBAC.

Item Writing

AIR will use, to the extent possible, the SBAC item/task/stimulus writer/developer training materials to begin item writing. AIR will use this material to train our item writers, and the item writing will begin in the item development system. The system facilitates item writing by giving writers all the tools and resources they need to allow them to focus on the creative aspects of writing while the system handles the clerical tasks. Assignments will be given based on feedback from the focus groups and will center on the hard-to-measure constructs. The item writers will focus on developing technology-enhanced items/stimulus and performance tasks.

When drafting items in the item authoring system, writers will see the complete texts of the target Common Core State Standards and the complexity level on the left side of the screen. On the right side of the screen, the writers will see a type-specific item template that has sections for each item component (e.g., stem, answer choices). These templates offer all the common features of word processors (e.g., spell check, formatting) and provide structure to the writers.

Internal Reviews

AIR's internal content reviews will ensure that each item written for the SBAC cognitive labs and small-scale tryouts will measure its intended construct. After item writers draft and upload the material that they have written using all the approved SBAC guidelines, the internal AIR review process will begin. AIR's test developers will ensure that the items and related materials comply with SBAC guidelines for clarity, style, accuracy, and appropriateness for the population being assessed. They will also consider the outcomes of the focus groups and ensure that the different formats/styles/innovations that are being questioned in the cognitive labs are present in different versions of the items.

For all items, the reviewers will verify that the item is properly aligned with the Common Core State Standard and SBAC item/task specifications. Based on the meticulous review of each item, the reviewers will accept the item as written, or will revise the item, attributes, or classification—or all three. A review alternative is to reject the item because it is too problematic in content, does not align with any standard, will not accomplish what is intended from the cognitive labs or small-scale tryouts, or all three. Whatever the recommendation may be, the review comments will be noted in the item development system. If an item is revised, its previous version will be automatically archived. AIR will review all items to ensure that they are clear and concise so that the focus can be on the aforementioned research goals of the cognitive labs and small-scale trials.

External Reviews

AIR suggests that SBAC review 20 percent to 25 percent of the items intended for use in the cognitive labs and small-scale tryouts. We propose preparing these items in batches, considering the outcomes from the focus groups and, later on, the cognitive labs. The initial batches will contain the selected sample items from the item/task and stimulus specifications. After the initial batch delivery, the items will be newly developed with input from the focus groups and then will be entered by the cognitive labs. SBAC will be able to review the items directly online in the item development system. During these reviews, SBAC will view the items along with the stimulus and graphics, using the same rendering engines that can display the items on the client-side testing systems.

Plan and Conduct Cognitive Labs of New Stimulus Types, Item Types, and Performance Tasks

Design and Protocols

We anticipate that we will recruit approximately 145 students per grade to participate in the cognitive labs, with the goal of having 10 students respond to each item. Approximately 60 students per grade will respond to a single performance task (three reading and three mathematics performance tasks per grade). The remainder of the students will respond to a mix of eight other items. Each cognitive lab interview will take 60 to 90 minutes, which in our experience is an ideal length of time for getting the maximum information from students without tiring them so that they are unable to meaningfully focus on the task. Students will be paid $60 ($50 for their participation plus $10 to cover transportation expenses). AIR will develop the protocols for the cognitive lab interviews and will train the interviewers.

The partners participating in this proposal have offices in a variety of states. To keep the costs of the cognitive labs reasonable, we anticipate that the majority of the cognitive lab interviews will be conducted in states where we have offices:

- California, the District of Columbia, Maryland, North Carolina, Oregon, Illinois, and Hawaii: AIR
- California: CTB
- Minnesota: DRC
- New York: College Board

We recognize that these are not all SBAC states, but the rapid timelines suggest that it is better to have the cognitive lab activities up and running as quickly as possible. AIR and our partners are prepared to conduct the studies in geographically diverse areas and will emphasize SBAC states where possible.

Because most of these offices are located in urban and suburban areas, we will make a special effort to ensure that students from rural areas are represented in the sample. There are rural areas within easy traveling distance of our offices in California, Oregon, North Carolina, and Hawaii, and we will locate space in those areas for several days so that we have interview sites that rural students can easily reach. We will also ask SBAC states with large rural populations to suggest additional sites for cognitive lab interviews. We discuss recruitment for the cognitive labs in more detail in Section 3.4.2.

Training

Training sessions for the cognitive lab interviewers will last approximately one day. Multiple training sessions will be held at each interview location to keep the number of participants in each training session small and allow individual feedback from the trainers. The training sessions will start with an overview to familiarize staff with cognitive lab principles and techniques. Participants will be shown a videotape of a cognitive lab interview so that they understand the process. Participants will then be paired off to take turns being the interviewer and the student and to conduct actual interviews with each other. Experienced cognitive lab staff will circulate and observe these interviews, stopping to offer critiques and suggestions for improvement.

Training is an iterative process that does not stop just because an interviewer has mastered the basics. We will also videotape (with the permission of the student and the student’s parent or guardian) some of the early cognitive lab interviews, and an experienced interviewer/supervisor will watch a tape of one of the first labs conducted by each interviewer. The supervisor will provide feedback to the interviewer. If the supervisor has any serious concerns about how the interview was conducted, he or she will observe the interviewer's next session either in person or on videotape to ensure that the interviewer is following the protocol and obtaining the needed information.

Protocols

Each cognitive lab protocol will start with an exercise to teach students to "think aloud." For example, the interviewer may ask a general question such as "How many students are in your mathematics class?" and then model how to think aloud while responding to the question: "There are five rows of desks in my math class and each row has six desks, so that is 30 desks, but two of the desks aren't used, so there must be 28 students in the class." The student will then be asked the same or a similar question and asked to think aloud while responding to it. The interviewer will encourage the student to continue to use this think aloud process while responding to the tasks presented in the cognitive lab.

Once the interviewer is confident that the student understands the process of thinking aloud, the interviewer will move on to the actual tasks to be assessed. The student will be asked to work through each task, thinking aloud as he or she completes it. If a student does not express his or her thoughts out loud, the interview will prompt him or her to do so.

AIR will develop a series of debriefing questions for students for each type of question evaluated in the cognitive labs (performance tasks, drag and drop, hot spot, etc.), and the interviewer will administer the questions after the student completes each test item. Debriefing questions will focus on identifying whether or not the student understood the question, and on identifying any parts of the question the student found difficult or confusing. For example, a debriefing question to elicit information on whether the student understood the task might ask, "Can you tell me in your own words what you think this question was asking you to do?" or "Can you explain to me why you chose this answer?" Interviewers will be required to ask all the debriefing questions, but they will also be allowed to probe further if they think that a student's response to a question was incomplete or if the student said something they think requires further clarification. Interviewers will not be allowed to interrupt a student to ask questions while the student is answering a test question, other than to remind the student to think aloud, because we want to evaluate how the student approaches each question without any outside interference.

Recruitment

It is important that the sample of students recruited for the cognitive labs be diverse and representative of the demographics of students who will participate in the actual test. To recruit as broad a sample as possible, we will advertise the cognitive labs through multiple venues. In recent years we have successfully advertised cognitive lab opportunities on Craig's List, so we will post a notice on the local Craig's List for the areas where we are interviewing. However, not everyone has access to the Internet, so we do not want to limit our recruiting to people who answer an online advertisement. We will also develop flyers advertising the cognitive labs and distribute those flyers where students are likely to congregate, including Boys and Girls Clubs, athletic facilities, libraries, community centers, and places of worship. If possible, we will provide flyers to local schools with diverse populations of students and ask them to send the flyers home with their students. However, in the past we have found that a school district's central office often needs a three- to six-month lead time to approve recruiting students for a project such as this. The timeline of this project will not allow that much time before we begin the recruitment process. We will work with SBAC states where we intend to do interviewing to see whether there is any way to get this permission more quickly.

The flyers and other recruitment materials will include a brief description of the cognitive labs and a local or toll-free phone number for interested people to call. In most cases we expect that a parent will make the phone call, but some students in the higher grades may call themselves. When we receive a call, we will give the caller additional information about the project. If the caller is still interested in participating, we will ask a series of brief questions to screen the student so that we can make sure that we are recruiting a diverse sample. We will work with SBAC to finalize the demographic characteristics of interest, but we anticipate that at a minimum we will ask screening questions to obtain the following information:

- Grade
- Gender
- Race/ethnicity
- Language spoken in the home (English, Spanish, other)
- Socioeconomic status (obtained by asking about either family income in broad bands (e.g., under $25,000, $25,000 to $50,000) or participation in the free/reduced-price lunch program)
- Type of school (public, private, parochial, home school)
- Whether the student has an IEP or 504 plan and, if so, what specific disability the plan addresses.

We may be able to schedule some students for interviews during the initial screening phone calls, but in many cases we will need to obtain contact information and then call students to finalize interview times.

AIR's Institutional Review Board (IRB) will review all recruitment materials to ensure that they are compliant with human subject research requirements. We expect that the IRB will require a signed permission form from a parent or guardian for all students under the age of 18 who participate in the cognitive labs. At the time a student is scheduled for an interview, we will email the parent/guardian a copy of the permission form (if the parent/guardian has access to email) and explain that the signed form must be brought to the interview. If the parent/guardian does not have access to email, we will either mail the permission form to the student's home or emphasize to the parent/guardian that he or she must be present at the beginning of the interview to sign the permission form.

Assignment of tasks to students
Prior to the start of the cognitive labs, the lead item developers will divide all the questions to be evaluated in the labs, except the performance tasks, into groups of approximately eight questions each. The groups will be constructed so that each includes different types of questions. We will also try to make the time students are likely to require to work through each group of questions approximately equal. Because students assigned to a performance task will get only one task, there is no need to divide the performance tasks into groups.

When a student is scheduled for a cognitive lab interview, the cognitive lab management team will assign the student to a specific set of questions or to a performance task. We will distribute the students across the tasks so that each task is evaluated by students with a variety of demographic characteristics as determined by the screening protocol described above.

Data analysis
Cognitive lab results are intended to be analyzed qualitatively rather than quantitatively. The cognitive lab interviews can help us identify types of questions or aspects of questions that may cause students difficulties that are unrelated to the construct being measured. The cognitive lab think-aloud protocol and the debriefing questions asked after students complete each task will help us identify these types of issues. Each interviewer will complete a standard template at the end of each interview to summarize the results of the interview. Interviewers will be instructed to complete the template immediately after the interview while the student's responses are still fresh in their mind, and to submit the completed templates to the project management team for review twice a week.

Some groups of questions will be designated for administration during the first week of cognitive labs, while other groups will be held for later in the process. At the end of the first week, the cognitive lab management staff will evaluate the results obtained thus far to see whether any item types appear to cause difficulties for students with specific demographic characteristics. For example, one or more item types may appear to be problematic for students who come from homes where a language other than English is spoken. If we observe this pattern, we will review the notes from the interviews with these students to try to identify which characteristics of the items make them more difficult for the particular demographic group. We will then work with the item developers to modify the remaining items of this type and use the later cognitive lab interviews to assess the results of that modification. In some cases, we may need to increase our recruitment quota for students from a demographic group to make sure that we are able to adequately evaluate the revisions made to the items.

We will continue with the cognitive labs, following this iterative process. The sample size for each item tested in the cognitive labs is not large enough to evaluate with certainty whether or not an item type exhibits differential item functioning (DIF). However, if the results of the cognitive lab interviews indicate that an aspect of a particular item type is causing students difficulty unrelated to the construct, we will attempt to modify items of the same type that have not yet been administered to fix the issue. As discussed below, after the completion of the cognitive labs we will evaluate the various item types in small-scale trials where we will have a large enough sample to measure DIF. The questions we pursue in our analysis plan for the small-scale trials will be informed by the issues that are encountered in the cognitive lab interviews.

Feedback

Information from the cognitive labs will feed quickly back into the item development process. The results of the cognitive labs will feed into the item development for the small-scale pilots, and findings will inform the item development for the pilot tests.

After all the cognitive labs have been completed, AIR will produce a report on the cognitive labs that at a minimum includes the following:

- Number and type of questions administered
- Demographic characteristics of the cognitive lab participants
- Key findings related to specific item types
- Efforts made to modify items based upon these findings
- Results of interviews conducted after modifications were made
- Research questions identified during the cognitive lab interviews that should be further explored during the small-scale trials.

Small-Scale Trials

Design and Protocols

The small-scale trials will be conducted in SBAC member states. Each SBAC state will be asked to submit a list of schools and districts in the state that includes demographic and other information that is readily available, such as average student reading and mathematics scores on current state tests, percentage of students in various racial/ethnic groups, and percentage of students receiving free or reduced-price lunch. AIR will work with the psychometric contractor to develop a sampling plan and choose a stratified random sample for the small-scale trials that is based on the information supplied by the states.

We will construct 12 fixed forms for each grade. The 12 fixed forms will be divided into three groups with four forms in each group. Each group of 4 forms will include 12 sets of six items, with items from each set distributed across the four forms (two items from each set of items on each of the four forms). This will result in each student's being presented with 24 items. Based on past experience, we think that students should be able to complete a test of this length in one to two hours.

In addition, we propose to solicit teacher ratings about each student's facility with the skills and knowledge in the content specifications. Specifically, for each claim, we will ask the teacher to rate the extent to which each student has achieved the skills claimed. These teacher ratings will provide an independent measure to be cross-validated with the measures from the assessment.

Recruitment

AIR will draft a letter to be distributed to all schools selected for the small-scale trials, explaining the purpose of the trial and what each school is being requested to do. AIR will consult with SBAC to determine whether the letter should be sent from SBAC, from the state superintendent in each state, or from AIR. AIR will include a toll-free number in the letter that schools can call if they have any questions. AIR will include a form with the letter for schools to fill out, indicating their willingness to participate in the small-scale trials and also designating a contact person (probably the school test coordinator) for the trials. AIR will follow up with all schools that do not return the letter. If we are unable to persuade a selected school to participate in the trials, we may ask SBAC or a representative from the state's assessment office to follow up with the school. However, AIR will make several efforts to contact the school and gain its participation before asking for help from SBAC or the state.

Participating schools will need to install the AIR secure browser on the student computers that will be used for testing, and train their staff members to administer the online tests. AIR will put together a manual explaining how to do this and will also build an online training module for test administrators so that they can be certified to administer the test. We anticipate that the online training will take a maximum of 30 minutes. SBAC member states that already use the AIR secure browser for state tests (Oregon, Delaware, Hawaii) will not have to install a new version of the secure browser; AIR will provide a link to allow them to redirect the secure browser from their state tests to the SBAC small-scale trials test.

AIR will provide a toll-free Help Desk phone number for schools participating in the small-scale trials to use if they encounter any issues either when preparing for testing or during testing. AIR currently staffs our Help Desk from 5 a.m. to 10 p.m. Eastern time, to provide support for states in different time zones. AIR will work with SBAC to determine when, within these hours, we will have support available for the states participating in the SBAC small-scale trails.

Assignment of Tasks to Students

Participating schools will be asked to send AIR a list of all their students in the grade(s) eligible for the study, along with the students' SSIDs, so that we can create identities for these students within the system to allow them to log in and test. We will request that the schools send demographic information for these students (race/ethnicity, gender, any disabilities, etc.) so that we can do a DIF analysis on the results of the small-scale trials as discussed below.

It is AIR's experience that it is easier for schools if all students at a given grade participate in the trial. In most cases, such clustering reduces the power of a sample. However, our proposed design leverages the clustering in a powerful way. Our research questions are designed around differences in the functioning of items, and the items will be administered in pairs. Where differences are of interest, the statistics become more precise in a clustered sample. Recall that the variance of a difference is given by var(a)+var(b)-2cov(a,b)—the covariance is subtracted off. If we are interested in the difference in the performance of items administered to the same student within the same school, the higher covariance term due to clustering provides an advantage! Somewhat paradoxically, while clustered samples yield relatively imprecise estimates of, for example, the percentage of students who get an item right, they can prove quite powerful in estimating the difference in the percent of students who get items a and b correct.

Data Analysis

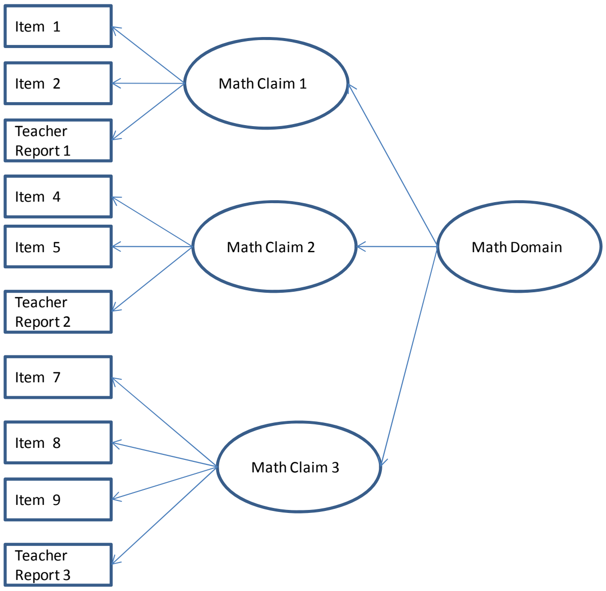
As discussed at the outset, the small-scale trials have three objectives:

- To provide student responses on which to train, evaluate, and refine scoring engines
- To test or confirm the findings from the cognitive labs in larger more diverse samples
- To systematically evaluate the impact of technology enhancements on the validity and fairness of the tests.

The scoring of the student responses to items and tasks is summarized in the next section. The analysis of those scores, and their use in training, evaluating, and refining the scoring engines, are described under proposal Part 2.

We propose to address the other two questions through a mix of simple analyses and more sophisticated structural equation analysis. The structural equation analyses will all build on the same framework. That framework is a second-order latent trait model, in which items load on the claims for which they are intended to serve as evidence, and those claims load on a common subject-area trait. This is schematically illustrated in Figure 19.

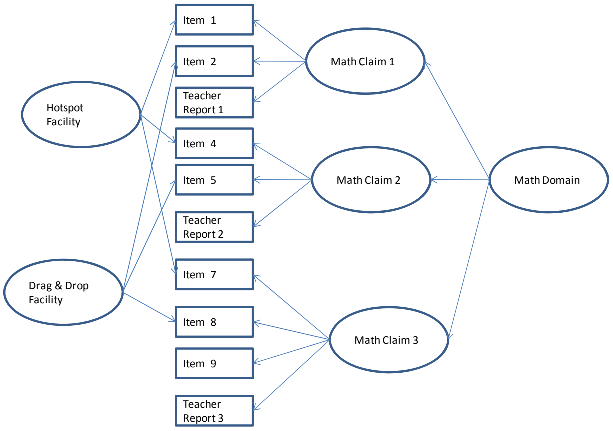
Figure 19: Schematic Illustration of SEM Framework for Study



The first and most basic finding that we hope to achieve from the cognitive labs is that performance on the items serves as evidence for the claims. This relationship can be tested within this framework. Using the same framework, it is also possible to evaluate whether the claims themselves have any divergent validity. A formal test of this would fix the loadings between first- and second-order traits at one, and evaluate whether removing that restriction led to significantly better model fit.

Comparisons of different interaction modes, item types, presentation layouts, and stimulus formats can be directly tested by introducing additional “nuisance” factors, as depicted in Figure 20. Here, we formally test whether the loadings on the newly introduced mode or format factors are non-zero. Non-zero findings indicate that the mode or presentation is differentially advantaging some students and not others.

Figure 20: SEM Model Augmented to Include Sample Construct-irrelevant Factors



These sophisticated analyses will target systematic differences across students' performances associated with different presentation and interaction factors. Much simpler analyses are appropriate to see whether some approaches generally remove (or construct) barriers to performance. We propose to compare scores on items of each type with scores achieved by overlapping sets of students on items of other types designed to measure the same or similar content. Systematic differences across modes, presentations, or other features suggest that the features associated with the less-difficult item variants universally reduce barriers relative to the other modes.

To evaluate fairness across identifiable groups, we will compare the number of items with each feature that are flagged for DIF. For this study we propose to use the standard Mantel-Haenszel procedure (Holland, 1985; Holland & Thayer, 1988) for dichotomous items, and the standard mean difference method (SMD) for polytomous items. Using flagging criteria negotiated with SBAC, we will investigate whether items with some modes, formats, presentations, or other features are more or less likely to be flagged for DIF.

SBAC is explicitly interested in one specific question: the effect of mixing item types and media on a single form. As described in the illustrative item development plan above, we plan to construct forms within each grade and subject to support this study. Some forms will be homogeneous, including limited and controlled item and media types. Simple comparisons across common populations tested with the different types of forms will evaluate whether student performance appears affected by mixing item types on the same form.

The analyses described here are illustrative. Until the final specifications are done and the hypotheses honed by cognitive labs, we expect that these plans will continue to evolve.

Scoring of Responses to Constructed-response Items and Tasks

Pilot items will be scored using automated scoring, human scoring, or a hybrid of both methods. One purpose of the small-scale trial will be to determine which scoring method works best for various types of items prior to the larger pilots. To determine this, we will need to collect and hand score student responses. These human-scored responses will inform the automated scoring system. After autoscoring, we can compare human and automated scoring.

The small-scale trials will also allow us to adjust the scoring rules prior to or during the human-scoring process. The goal is to develop an item pool that is, for the most part, autoscored. However, some items may lend themselves better to human scoring. Such items will need to have the most appropriate rubrics associated with them before they become part of the pilot pool. Below we describe a study developed to evaluate and compare different rubric styles.

DRC will lead the human-scoring efforts. DRC brings a tremendous amount of experience scoring constructed responses with 24 years of experience in delivering accurate scores for millions of students in numerous client states, including Alabama, Alaska, Arkansas, Florida, Kentucky, Louisiana, Minnesota, Nebraska, New Jersey, North Carolina, Ohio, Oklahoma, Pennsylvania, South Carolina, and Washington.

Our proposed plan for performance scoring of the constructed-response items begins with the sample size. Approximately 36 constructed-response items per grade and content area will be used during the trial for a total of just over 500 unique items. For each of the 500 items, we will collect a sample of approximately 200 responses. Each response will be scored by two independent readers. Any responses with nonadjacent scores will be read a third time for resolution by an expert scorer. DRC realizes that these scores must be accurate and that they will provide a reference point for automated scoring during the pilot test. DRC's training and scoring processes are ISO 9001:2008-certified and have robust quality checks in place. These scoring processes are described later in this section.

Performance tasks will also be administered during the small-scale trials. Using 32 performance tasks across subjects, we will sample approximately 150 students per task. Those tasks will be scored by DRC's professional scoring population in the fall after rangefinding has been conducted. Each response will be read one time with a portion read a second time as a measure of rater reliability. A subset of these items will be used to study rubric models and to determine the most appropriate method for conveying scoring rules.

An initial rubric will be written for each item when the item is initially developed. The test development team will also be on the lookout for items that might lend themselves to more than one rubric style. Any such items will form a subset of items that will undergo additional scoring using the additional unique rubrics. We have initially identified the following rubric styles for potential use in the study: holistic, generic, item-specific, checklist, analytic, and weighted analytic. Items identified for inclusion in the subset sample will be scored multiple times by different teams of scorers using two or more unique rubrics. The resulting scores will make it possible to compare different ways of structuring scoring rules for human use. We can also use this information to determine which rubrics generate the most comparable autoscoring results. All this information will be available to the Psychometric Services provider as we work together to determine criteria for automated scoring adequacy to present to SBAC.

DRC's Performance Assessment Services team understands that the proposed work will require careful planning, thorough and thoughtful system designs, and sound execution of established procedures. DRC's Content Specialists have many years of experience with large-scale high-stakes assessments, and their management lead has firsthand experience monitoring similar projects, working with multiple entities, and producing accurate results. We are confident that DRC's knowledgeable and experienced staff will successfully score the small-scale trial and that SBAC will be pleased with the results.

Scorers

DRC is able to tailor its scorer staff to the client's program. All scorers will have a minimum of a four-year college degree with a background in the content areas being assessed. Preference will be given to candidates with previous experience scoring large-scale assessments and with degrees emphasizing the appropriate content areas. During a personal interview, scorer candidates must respond to a DRC writing topic to demonstrate writing proficiency. Mathematics scorer candidates must also successfully solve a series of DRC mathematics problems and show all steps necessary to reach the correct answer.

Image Handscoring

Now in its tenth year of operation, DRC's Image Handscoring System has proved to be highly efficient and completely accurate for scoring large-scale assessments. The dynamic system allows scorers to score items online, increasing efficiency by eliminating the routing of paper and eliminating the possibility of lost student answer documents. Instead, student responses are electronically routed to geographically dispersed DRC Scoring Centers. Responses are allocated to scorers through a custom dealer program, ensuring that each scorer is assigned a random workload that allows the project to be processed in the most efficient manner.

The system provides the scorers with the ability to view full-page images from multiple perspectives, such as zooming in/out and image flipping or rotation. Images remain intact with the various viewing capabilities and cannot be modified by the scorers.

The Image Handscoring functionality applies a set of process rules and client-defined read-behind criteria (i.e., all responses receive two independent reads and responses with nonadjacent scores receive an independent third reading). The programmatic Item Definition Application defines the possible score values, the possible nonscore values, and the applicable scoring rule.

The Image Handscoring functionality requires scorers to forward all potential nonscorable responses to the Scoring Director. Only the Scoring Director is able to assign the nonscorable code. Scorers "alert" any responses that indicate potential issues related to the student's safety or to potential plagiarism.

Each handscoring site is connected to the main DRC operations facility through a DS3, 45 Mb/S private circuit. The operations facility has secure database servers and multiple applications that support the scanning, editing, scoring, and handscoring processes. Database backups and archived images are stored off-site on tape media for disaster recovery purposes. Each DRC scoring site has a server and a local area network (LAN). Scorers, Team Leaders, and Scoring Directors connect to the LAN via hundreds of PC workstations, and use locally resident software to view and score student responses. Authorized on-site DRC personnel (e.g., Content Specialists, Project Managers) can access the LAN to recall student responses.

DRC's Software Quality Assurance Analysts test the imaging system to verify that all handscoring programs are compliant and in place for performance assessment personnel prior to the transfer of production images. Images produced from test files are generated, processed through the handscoring system software, and accessed from the handscoring client software, where Quality Assurance Analysts score the test images using the handscoring criteria and specifications just as they will be in the production systems. Throughout this testing cycle, multiple quality checks are executed to ensure that the data integrity for each student record is intact and accurately reflected in the scoring database.

Handscoring Procedures

DRC's high-quality training, scoring, and monitoring processes have been used for years to handscore student responses. DRC will uphold its dedication to accuracy and quality under this new contract.

General Scoring Procedures

- Readers are seated at imaging stations and are assigned unique ID numbers and passwords.
- The Scoring Director provides detailed directions for using DRC's computerized handscoring system.
- Student responses are separated by item and routed to qualified scorers. Readers cannot tell whether they are conducting first or second readings. They can forward responses to Team Leaders for assistance.
- The process of routing and scoring responses continues until all responses have received the prescribed level of readings.
- Ongoing quality-control checks and procedures monitor and maintain the quality of the scoring sessions. If any unusual data are observed, DRC investigates and resolves any issues.
- Responses can be retrieved on demand (e.g., specific batch files, specific grades, specific students), should the need arise during or subsequent to the handscoring process.
- Should the need arise, responses can be rescored based on item- or response-level information, including item number, date, score value assigned, or reader ID.

Monitoring and Maintaining Handscoring Quality Control

Accurate and consistent results are the backbone of all handscoring activities. The following methods ensure that DRC will provide accurate and reliable scores:

Anchors, which are prescored student responses, will define and exemplify the score scale. Scoring Directors and Team Leaders will lead a thorough review of the anchors along with the scoring rubrics to ensure that scorers understand how the rubrics should be applied to student responses.

During an intensive training session, scorers will practice and demonstrate scoring proficiency by accurately scoring sets of training responses. Scoring Directors and Team Leaders will review the training responses with the scorers to ensure that they understand how to score the responses.

Qualifying responses are similar to training examples in that they have been prescored through rangefinding. The responses will be divided into sets and scored independently by each scorer trainee. The data from these qualifying rounds will be used to determine which scorer trainees will be qualified for actual scoring.

Team Leaders will conduct routine read-behinds for all scorers. DRC's imaging system allows Team Leaders to determine read-behind rates (frequency of monitoring) for each scorer. The imaging system randomly selects which responses the Team Leader will read behind.

To monitor scorer reliability and maintain an acceptable level of scoring accuracy, DRC will closely review inter-rater reliability and score point distribution reports that will be produced daily.

Scoring Directors will meet regularly with their Team Leaders to review the quality control statistics. If a scorer falls below acceptable agreement rates, the Team Leader or the Scoring Director will retrain the scorer and, if necessary, remove all assigned scores given by the scorer in question. The images will then be redealt and rescored.

Quality-Control Reports

Report	Report Specifics
Inter-rater Reliability	Monitors how often scorers are in exact agreement and ensures that an acceptable agreement rate is maintained. Provides daily and cumulative exact and adjacent inter-scorer agreement and the percentage of responses requiring resolution.
Score Point Distribution	Monitors the percentage of responses given each of the score points. For example, for a four-point constructed-response item, this daily and cumulative report shows how many 0s, 1s, 2s, 3s, and 4s a scorer has given to all the responses he or she has scored prior to the time the report is produced. It also indicates the number of responses read by each scorer, so that production rates can be monitored.
Item Status	Monitors handscoring progress. Tracks each response and indicates the status (e.g., "needs a second reading," "complete") to ensure that all responses are completed.
Responses Read by Scorer	Identifies all responses scored by an individual scorer. This report is useful if any responses need rescoring due to potential scorer drift.
Read-Behind Log	Is used by the Team Leader/Scoring Director to monitor intra-rater reliability. Team Leaders read a random selection of scored responses from each team member. If the Team Leader disagrees with the scorer's score, remediation occurs. This has proven to be a very effective type of feedback because it is done with items live-scored by a particular scorer.

Part 4 Study of Item Procurement Options

L1 – 11 Review and evaluate the processes for three proposed procurement options (state submitted, state managed, SBAC managed) for generating constructed response and selected response items (with associated stimuli) that align with SBAC item/task and stimulus specifications.

Determine the effectiveness of all item procurement options. In a small, controlled way, try out state managed, state submitted, and SBAC managed item/task development protocols.

The Collaborative is pleased to present a plan for an independent study to assist SBAC in evaluating different item procurement options. Leveraging various procurement methods and existing assessment content can be an important factor in the cost-effective development of a balanced assessment system. The work of this task will be primarily conducted by CTB in partnership with HumRRO, an organization with extensive experience in conducting independent evaluation studies.

HumRRO currently serves as independent evaluator for several important efforts, including the National Assessment of Educational Progress (NAEP). For the NAEP program, HumRRO staff frequently monitors item development, test administration, scoring, and analyses and reporting processes by observing first-hand the activities associated with implementation. This involves observing the training of contractor staff, followed by developing protocols and other data collection materials. HumRRO staff then use these materials when observing implementation of the processes to gauge the extent to which contractor staff complete tasks as they were trained to do, administer the processes in a standard manner, interpret and apply rubrics as intended, maintain security of items and other test materials, report results appropriately, etc. We also employ these materials when reviewing the contractors' quality control plans to identify steps within the various processes that are ambiguous, do not pinpoint the responsible party, or do not ascertain how handoffs will occur between multiple contractors. Based on policy and agreed upon practices, HumRRO staff use their understanding, expertise, and observations to formulate recommendations to strengthen procedures. We routinely conduct special studies for the

NAEP program to investigate areas of particular challenge, including a recent effort that examined the quality of NAEP mathematics items. The role that HumRRO plays on the NAEP program and the expertise staff has developed from conducting evaluations of other education programs are directly relevant to the current effort.

For the purpose of this proposal, HumRRO and CTB have made the following preliminary assumptions about the study. These assumptions will be reviewed and revised in consultation with SBAC before the study parameters are defined and the study begins.

- SBAC-managed option: SBAC-managed items are defined as those that are being developed under this contract. Items for evaluation of this option will be selected according to the study parameters from the approximately 10,000 items developed under Task 1 of this contract.
- State-managed option: State-managed items are those that are developed specifically for SBAC by states. Items for evaluation will be selected from pools of items developed by states using SBAC specifications, training materials, etc. The costs of this development, i.e., stipends for teachers and others involved in authoring and reviewing items, are included in this proposal for two states participating in this option.
- State-submitted option: State-submitted items are those that are part of existing state-owned item banks, are under development by states for their current assessments, or are otherwise available to states to administer. Items will be reviewed and aligned with the SBAC item specifications, content specifications, the CCSS, or other appropriate frameworks. Stipends for teachers or others to carry out such alignment are included in this proposal for two states participating in this option.

We propose the following activities as part of this evaluation study.

Definition of Implementation Parameters

An early activity related to the study of item procurement options will be to refine the definitions of the three options as well as to clarify and/or solidify the parameters for state participation, especially for the state-managed and state-submitted options. To ensure results from this study allow for a fair comparison of the options, we will work closely with key SBAC staff to delineate and refine (as appropriate) the definitions of each procurement option and the parameters for state participation in each.

At this time, for the state-managed option, we believe that it will be important to establish the extent to which states must follow common guidelines for recruiting, training, and monitoring item writers, and for editing, reviewing, and revising items once they are developed. We assume that states which participate in this option will use the specifications, training materials, etc., that are produced by SBAC.

For the state-submitted option, guidelines for developing constructed-response and selected-response items will need to be established. At this time, we define this option to mean that states will either manage or contract for item writing, with review and revision completed centrally by SBAC. This is in contrast to the state-managed option, where review and revision will be handled by the individual states. Even under the SBAC-managed option, states that participate must allow access for educators to participate in the common writing, editing, and review processes. The participation policies developed by the SBAC 08 will apply to this option as well as to the state-managed option.

Sampling Plan Design

We recommend the use of random assignment when conducting the small-scale study to compare procurement options. Unless there is some element of random assignment to the procurement options, we recognize it will be very difficult to control for differences across LEAs and states, which would limit the generalizability of the evaluation results. We suggest that sampling occur at the state level. We believe it is unlikely that a state could participate in both the state-managed and state—submitted strategies. Even if a state could manage complete development of some items while submitting others for external review and revision, the likelihood of cross-contamination would undermine the desired comparisons. We expect that most, if not all, of the participating states could be included in the SBAC managed option (e.g., supplying developers or reviewers to work under the direction of SBAC staff and vendors).

How many states? Although SBAC estimates that up to two states will participate in each of the state-managed and state-submitted options, we strongly recommend recruiting at least 15 states to participate in the study and randomly assigning five states to each of the three procurement strategies. We recommend this larger sample of states for two reasons. (a) Including more states in the study will allow for a sufficient number of items to be reviewed for quality and cost (without burdening only a few states to supply these items), and (b) there will be more confidence in the results. To the extent possible, sampling should be stratified on important state differences, such as size, so that an equal number of states within each stratum is assigned to each strategy. A second option would be to randomly assign available states to the first two options and then use developers and reviewers from the remaining states in the SBAC managed option. The sampling of states will necessarily be more complex, if some states cannot participate in the development of all item types.

How many items? The next step in the sampling plan will be to identify the number of items of each type that should be produced under each procurement option. If at least 15 states can be recruited to participate in the study, we propose a minimum of 500 items be developed or submitted per state (divided across grades, subjects, and item types). However, if only two states can be recruited, then each state would need to develop or submit 1200-1500 items. From the overall pool of items, we will randomly select the designated number of items (taking care to select an approximately equal number of items from each state). We expect this approach will prevent "cherry picking" of the best items by any state while allowing for a manageable number of items to be evaluated.

Implementation Timeline

The timeline will necessarily be aggressive. We suggest including items from each option in the 2012-13 Pilot Test so that pilot test data, along with results from independent item reviews, can be used when evaluating items developed under each option. Sample milestones for the evaluation include the following:

Project begins	April 2012
States recruited for/assigned to the three procurement options	May 2012
Items developed/submitted	May – September 2012
Independent item reviews conducted	September – December 2012
Pilot test	January 2013 – June 2013
Evaluation analyses and reporting	June– August 2013

Evaluation Methodology

The evaluation methodology will include procedures to document the item development processes implemented under each option, and identify outcome measures of cost and quality. We will collect as much data on implementation as possible by attending or observing pre-planned committee meetings. However, it may be necessary for us to extend those meetings (or possibly convene additional ones) to collect supplemental data to inform the comparison of the procurement options or to verify states' claims of implementation procedures. Minimally, we plan to have a representative participate in the online meetings that involve item bias and sensitivity reviews as well as participate/observe in a sample of the online item development meetings.

For outcome measures, HumRRO will examine the time and costs required to develop items under each option, including the time and costs associated with item revision and refinement, and the resulting quality of the items that are developed. Costs for state-managed and state-submitted options must include time required for state department staff as well as for item developers, editors, and reviewers. For each participating state, we will seek to identify a maximum rate of item development. Given that funds will be provided from this contract for the states participating in the state-managed option, an important consideration will be the ability of the states to sustain development of quality items.

Until field test data are available, we must rely on the judgment of experts to determine the quality of the items developed/submitted. To assess item quality, we expect the experts will rate the items on a number of item quality parameters, including:

- Specific match to Common Core State Standards/Blueprint
- Cognitive rigor/item depth of knowledge (DOK)—(this and above parameter used for alignment review)
- Item content
- Item craftsmanship.

We propose to randomly select items from each of the three procurement options and provide them to the experts to rate; we anticipate this will occur during one of the scheduled content reviews. To capture a measure of reliability, we will have a reasonable percentage of the items rated by at least two experts (about 75 percent). Further, we propose to include a fairly large number of items per assessment target (at least 15); estimating between six and 15 assessment targets, this means between 90 and 225 items per grade subject will need to be evaluated. Assuming one third of the items come from the state-submitted option, state-managed and the SBAC-managed option groups will need to generate 30-75 items per grade per subject. Proportions of item types (multiple choice, constructed response, technology enhanced) should mimic targets for the operational assessment.

Because the test will eventually be adaptive, we will not be able to produce complete alignment results without the algorithm in place and some test takers from which to generate sample “forms.” For this study, we will look at only whether there are sufficient items within each assessment target, with appropriate cognitive rigor/depth of knowledge, to generate forms with acceptable alignment.

Once field test data are available, we will examine classical item statistics and flag items whose functioning appears questionable. For constructed-response items, this will include review and evaluation of the scoring results. For selected-response items, HumRRO will examine the distractor options and identify any incorrect options that are selected by relatively high-scoring students as an indicator of possible keying errors or ambiguity in either the stem or distractors.

Evaluation

The evaluation report will describe evaluation procedures, outcomes, and recommendations that highlight strengths to continue, as well as improvements to enhance the process. The evaluation will include a formative component (that suggests improvements to each of the procurement options) and a summative component (that compares the resulting cost and quality of items developed under each procurement option).

B. Work Plan

We have created an integrated work plan that describes the major activities involved in each task and that highlights the interdependencies and relationships among the separate tasks of the proposal. For clarity, we have divided the work plan by task and will note interdependencies with other tasks and other vendor deliverables. The details of our processes and technical methodology have been discussed in the above sections. This section provides an outline of the planned work tasks with a brief summary of each activity. For additional details of the approach for each task, please refer to Sections L1-12, 18, 19, and 20.

Our Management Proposal provides details of how we will organize and manage the work of Tasks 1-4 as well as our communication plan. Development of the required reports of research and development are discussed within the appropriate tasks.

We have created an integrated work plan that describes the major activities involved in each task and that highlights the interdependencies and relationships among the separate tasks of the proposal. For clarity, we have divided the work plan by task and will note interdependencies with other tasks and other vendor deliverables. The details of our processes and technical methodology have been discussed in the above sections. This section provides an outline of the planned work tasks with a brief summary of each activity. For additional details of the approach for each task, please refer to the above sections.

Task 1 Oversight, Item/Task/Stimulus Development and Reviews

Our plan for the development of items for the Pilot Tests involves both Collaborative-developed and teacher-developed items. The Collaborative member staffs are well-versed in the CCSS and have studied the SBAC Content Specifications. We are currently reviewing the draft Item and Task Specifications released on Jan. 6, 2012. Our design and content experts will continue to review and analyze SBAC documents as they are released, independent of anticipated contract award dates. Given the tight timelines for the development of items for the Pilot Test, our team will begin planning and developing training strategies as documents become available. Should the Collaborative be identified as the apparently successful contractor in February, we will utilize our preliminary planning to finalize plans for using Item/Task Specifications, participation policies, training materials, style, bias/sensitivity, and accessibility guides, and other SBAC documents that will influence the development of stimuli, items and tasks for the Pilot Test. The resulting plan will be presented to SBAC immediately upon contract award in late March.

Activity 1: Development Planning Meeting

Upon contract award, we will organize an initial development planning meeting. This two-day, face-to-face meeting will include two representatives from SBAC and the development leads, subject-matter experts in ELA/Literacy and Mathematics, and other staff from the Collaborative. This meeting will focus on two topics. Prior to the meeting, our development teams will have reviewed the SBAC Item/Task Specifications, training materials, blueprints and other available documents. We will prepare a draft content development plan for review at the meeting that details the required number and types of items/tasks and stimuli for each grade and content area by type, assessment target, and cognitive complexity. This meeting will provide an opportunity to get important feedback from SBAC as to the overall requirements of the item pool and suggested revision to the content development plan. This plan will be finalized after the meeting and provided to SBAC for approval.

The second primary topic of the agenda will focus on understanding the item specifications and other requirements impacting the development of items and tasks for the Pilot Test. To begin the discussion, the Collaborative will present a summary of our review of the specifications and will identify topics for clarification with SBAC. We anticipate that this review of specifications will take approximately half a day. Following the specifications discussion, we will present a summary of our review of the training materials and our plan for implementing these materials within a trainer-of-trainer model. These discussions will be structured to allow other SBAC content area representatives to join by phone or webinar. After the discussion, we will document any collective decisions about using the specifications and training materials. The outcomes from this meeting will be specific training plans for teacher authors and vendor authors for stimulus, item and task development.

Activity 2: Stimulus Development

Given the constraints of the timeline for the development and review of stimuli, the Collaborative proposes to locate and develop stimuli using vendor resource. We will train stimulus writers to perform assigned stimulus development tasks. Training will be provided to ensure that these individuals are selecting or developing stimuli that reflect the spirit and requirements of the common core standards. Based on the draft list of specifications topics for ELA stimuli, we anticipate that the training materials developed by the SBAC 08 vendor will address general requirements of the specifications as well as more in-depth training in requirements by grade level and text type. We anticipate that training will also address requirements for text complexity, text features, and text structure by grade level and text type. Stimulus writers will participate in training that addresses general requirements as well as requirements that are common across grade levels, including instruction on how to use the online authoring system to submit stimuli and document metadata, including source documentation. We assume that in-depth training in developing stimulus materials that will support CCSS-aligned items will be organized by grade span. We recommend that as part of the training, writers be instructed to also create and submit text maps which will outline the theme or main ideas, craft and structure elements, and other key elements of the text that will support CCSS-aligned items.

This plan will bring diversity to the process of locating and authoring stimuli that meet the stimulus specifications. Stimulus development, including training for stimulus developers, will begin immediately upon contract award based on our overall understanding of the item pool requirements. Specific numbers of stimuli required will be adjusted once the content development plan has been finalized and approved as a result of the Development Planning Meeting.

Activity 3: Stimulus Reviewer Training and Stimulus Review

Due to the aggressive timeline for item development (particularly for the cog labs), stimulus reviewer training and stimulus review will overlap. We will work closely with SBAC to immediately implement the SBAC participation policies and identify a pool of qualified individuals for reviewing stimuli. These reviews will be conducted online and will be staggered with the development of stimulus materials. This will allow for the rolling review of stimulus materials as they are identified or created as well as provide reviewer input into the further development of stimuli. The schedule in Section C indicates how stimulus development is phased with item and task development.

Activity 4: Recruitment of Item Authors

Based on the participation policies and item writers identified by the SBAC 08 vendor, we will work closely with SBAC to finalize procedures for recruiting and contracting with item developers/writers. Proposed selections and qualifications of potential teacher item writers will be provided for SBAC review and approval prior to final contracting. Our Management Plan provides details of how we will contract and compensate individual authors.

Activity 5: Item Development Training

Our plan for teacher-authored item development is based on providing training and support for small groups of teachers who will work closely with an editor to develop a group of items targeting a single content area claim at a grade band. Each small group will include approximately three teachers and one vendor editor/facilitator with expertise at the appropriate grade band and content area. These groups will work closely to author, peer edit, and review items and tasks according to specifications for a specific claim. Editors and teachers will be trained using a train-the-trainer model as described below. Collaborative editors will be trained as facilitators and will then provide training to teacher authors.

Item development facilitator training

This online training will include Collaborative development and content leads, and the editors who will serve as facilitators for the small groups of teacher authors and others who will be managing the vendor-developed items. This training will use the SBAC Item Writer Training materials and will focus on the preparation of the participants to appropriately deliver training to item authors. It is anticipated that this online meeting would span several days to allow for both training and certification of facilitators. SBAC representatives will be invited to attend as participants or observers. Selected teachers (from the recruited poll) may be invited to participate at this level to provide feedback on implementing this training with the teacher authors.

Item Writer Training

Item Writer training will consist of two components. The first component will be a general training for all item authors using appropriate components from the SBAC Training Materials. This component will cover the CCSS, the SBAC Content Specifications, the SBAC Item/Task Specifications, descriptions of the various SBAC item types and technical instructions on using the vendor's item authoring and development tools. The second component of the training will be organized by grade band and content area and will address issues specific to each group. We anticipate that this training will be online and will be configured for a few hours each day over multiple days to allow for processing of information, individual and group practice, etc. While the same training material will be used for all authors, separate sessions will be held for teachers and for vendors to allow the training materials to be focused closely to the needs of the participants.

Activity 6: Item Development

After the Item Development Training, teacher authors and vendors will begin work. Items authored by vendors will proceed according to the internal development processes of the Collaborative members, described In Section L1-12. Teacher authoring will be done in facilitated small groups where the editor can work closely with teachers to complete item development assignments. The chart below shows our initial development plan across the Collaborative; this chart may be adjusted later as we have more information about the specific expertise needed to address specifications at individual grade bands and content areas.

Table 19: Collaborative Development Plan

Subject	Grade band	Grade Band Lead	SR/CR (conventional and TE)	Performance Tasks (conventional and TE)
ELA	3-5	DRC	CTB, AIR,	CTB
ELA	6-8*	CTB	AIR, DRC	CTB, CAE
ELA	HS*	CTB	AIR, DRC	CAE
Math	3-5	AIR	CTB, DRC	CTB
Math	6-8*	CTB	AIR, DRC	CTB, CAE
Math	HS*	CTB	AIR, DRC	CAE

*College Board CCR audit of items

The Collaborative editors will work closely with the teacher authors to produce items and tasks that will meet the purposes of the SBAC assessments. Once the initial group item writer training has occurred, individual groups of editor/teachers will begin work. We anticipate the following work plan for the authoring teams to include three phases.

- Phase 1: Team planning meeting for either a passage (ELA) or a claim/standard (math). During this meeting, the team will study the relevant item specifications, discuss how the standard(s) grow across the grades being considered, the range of content covered, what evidence needs to be elicited from students to support the claim, and general ideas for items/tasks…?
- Phase 2: Peer review of initial drafts of items. This session, facilitated by the editor, will be an opportunity to check the items against the specification and vision and the team's plan for the larger item set for the passage/standard.
- Phase 3: Weekly check-in sessions. The editor will schedule regular sessions with the team for check in. Some of these may be with other teams to look at the draft items as a collection. This is particularly important for ELA where the teams focusing on the different claims come together to review the entire set for a stimulus. These meetings can also resolve problems/issues and provide gather informal feedback on the process/training to date.

These three phases repeat when the team moves on to another stimulus or standard.

Activity 7: Preparation of items for Committee Review

The Collaborative will provide a review of all items (teacher-authored and vendor developed) prior to presentation to review committees. As a result of our innovative model for working with teacher authors, items will need little additional editing. In accordance with the response to Q8 and Q19 from the Q&A document, the Collaborative will review all items prior to committee review using best practice guidelines. All edits will be tracked in the interim item authoring system for later review by committees and SBAC.

Activity 8: Reviewer Training

Review committee members will be identified based on SBAC participation policies and the procedures described above. We propose a series of online trainings using the stimulus and item content review committee and facilitator training materials developed by the SBAC 08 vendor to train reviewers in how to use the Stimulus Content Review Guidelines or the Content Specifications and Item/Task Specifications to review test items for content alignment, depth of knowledge, and content accuracy and appropriateness. We will also train stimulus and item content reviewers in the use of established protocols for recording their responses to the content review of items and capturing their recommendations for item revisions.

The training sessions will include sufficient practice opportunities to determine that individual reviewers will apply the review checklists, etc., appropriately and will provide a comprehensive review of the items within their assigned area (content, bias, sensitivity, accessibility). Because these reviews are uniquely different, our editors will provide additional support to reviewers for the initial phases of the item review to be sure reviewers understand all guidelines and procedures.

We also anticipate the SBAC staff and/or work group members may wish to participate as additional members of the content, bias/sensitivity or accessibility review meetings. This participation can be easily accommodated within our authoring system, and SBAC comments can be incorporated into the pool of reviewer comments if so desired.

Activity 9: Item Reviews

Items reviews will be carried out as described above in Objective 12.The Collaborative will establish contracts with reviewers according to SBAC-approved protocols. There will be separate reviewers focused on content, bias and sensitivity, and accessibility. We propose that the three reviews be conducted simultaneously in addition to independently. Because reviewers will be located across the country, the reviews will be conducted online, using secure protocols. The interim item authoring system that we are proposing to use will manage secure, electronic workflows whereby items are assigned to participants for review. The system will track the status of review workflows as well as reviewers' individual item ratings and recommendations or comments. Record-keeping will be conducted according SBAC-approved mechanisms, with the support of the documentation and tracking features of the item authoring system.

Activity 10: Final review with SBAC

The Collaborative will assemble documentation of all reviewer comments and make it available for SBAC review. We will work with SBAC staff and appropriate work group members to plan the format and implementation of this review round. We will provide recommendations for edits to items based on SBAC procedures and protocols.

Activity 11: Final Revisions to Items

After consultation with review with SBAC, the Collaborative will apply approved revisions to items.

Any item or task revision will be made following the same protocols/criteria used during item development and item reviews as described in SBAC-08. After revising an item or task, it will be essential to check the final version again to confirm that it meets all item review criteria for content, bias and sensitivity, and accessibility (i.e., to ensure that the revisions effectively and appropriately resolved the stated issues and do not introduce any new issues that would affect the validity, reliability, fairness, appropriateness, or accessibility of the item or task). All final versions of the revised items will go through a final QA review by CTB's trained Editorial Quality Assurance reviewers. This final review will focus exclusively on performing a final style editing and copy editing of the items. At this point, accepted items have been deemed to meet the requirements of the item specifications.

Activity 12: Final Deliverables Upload

The final item pool of approved items will be provided to the SBAC 07 vendor for upload into the Item Authoring system according to the required interoperability criteria. The Collaborative will perform a final quality control check on a sample of items to confirm the item upload.

Task 2: Automated Scoring and Scoring Models

Our technical approach provides the details of our workplan, along with the rationale for the steps and their sequence. The steps and sequence deviate from the list of tasks included in the table on page 28 of the RFP, but cover all of those tasks. The reshuffling of tasks comes in response to the significant timeline risks, and our approach to mitigating those risks. To assist in the review of this proposal, we identify the proposed task that covers each of the tasks identified on page 28 of the RFP in Table 20.

Table 20: Required Activities

Task identified in RFP	Activity number under which work will be performed
Develop a vision for a comprehensive approach to score constructed response and technology-enhanced items using automated or hybrid scoring. Address performance tasks using automated, human, and hybrid scoring (see page 36).	Activity 4
Develop and implement recommendations for validation of item/task scores.	Activity 3
Develop and test multiple open source automated scoring mechanisms/programs (including innovative scoring options).	Activity 2, Activity 3, Activity 10
Obtain consultation, review, and approval of recommendations by SBAC Technical Advisory Committee.	Activity 11
Develop materials and processes for selecting range-finding examples and conducting range-finding.	Activity 7
Conduct range-finding activity to calibrate automated scoring engine.	Activity 8
Use small-scale trials to develop a model for use of range-finding to calibrate automated scoring engine for pilots and field tests.	Activity 6
Conduct studies validating the comparability of automated and human scoring for different types of constructed response and technology-enhanced items and performance tasks.	Activity 3
Meet all required inter-operability standards.	Activity 2

Below, we briefly summarize each task in the workplan. Please refer to the Project Approach for full details and the rationale.

Our project approach also proposes an additional activity, priced separately, to help establish an open-source organization to support and maintain the open-source scoring engines, and potentially, other open-source components developed by SBAC.

Activity 1: Gather and Prepare Existing Student Responses

In order to gain extra time to develop and hone open-source scoring engines AIR proposes beginning that process immediately. Therefore, a substantial amount of the work on the scoring engines can be accomplished before the data from the small-scale trials become available. To accomplish this, will ask SBAC and its member states to identify previous test items and tasks that approximate the types of items and tasks that SBAC would like to automatically score. We will ask SBAC leadership to assist in this effort. We will work with SBAC representatives and, drawing on our experience with this sort of transcription, draft a set of transcription rules that leaves no room for transcriber judgment.

Activity 2: Gather Open-Source Components

AIR has conducted an initial review of available open-source scoring engines and relevant scoring engine components that may contribute to this project. We will prepare a summary memorandum comparing the benefits for application to each item type for each relevant component, and work with SBAC to choose the open-source tools on which the engines will be built.

For each candidate open-source component, AIR will summarize its capabilities and identify a list of benefits and liabilities of the tools. We submit this summary to SBAC to inform a discussion to select the tools to be integrated into the open-source scoring solutions. AIR will prepare and deliver a detailed design document and unit test plan for the scorers that we intend to develop. As described above, the "testing" (discussed in greater detail below) will also evaluate the ability of the engine to self-report high-risk cases-those that may not be scored correctly.

AIR will then write the “executive” code needed to turn the open-source libraries into capable engines. The details of this work are described in the technical proposal. This task will result in technical specifications, design documents, and actual open-source software.

Activity 3: Conduct Initial Analysis

With the first versions of the scoring engines in hand we will begin to test them against the sample of items and responses gathered from the states. We will divide the student responses into two groups: one to use for training and a second, independent set to use for validation. AIR will work with SBAC to identify the outcome measures on which to judge success. In each case, we will compare the success of the scoring engines with human-scored counterparts.

Activity 4: Prepare Vision Document

Early in the process, AIR will work with SBAC to develop the Vision Document outlining the vision for the scoring system. The Vision Document will express the goals of a comprehensive scoring system. These goals will be related to the current state of the art, and the vision will set forth a realistic path from current capabilities to future aspirations. Definition of the current state of the art will reflect the combined industry experience of the CTB-led team, a review of the literatures, and initial findings from the preliminary analysis. AIR will then work with SBAC representatives to define this vision. The Vision Document will be concise, so that it may efficiently serve as a touchstone for decisions made during the requirements analysis and definition process.

Activity 5: Prepare Recommendations for Validation of Scoring Models

In this task we will seek to "validate" the machine-scores assigned to items. In our project approach we enumerate a draft list of statistical indicators that will be used to evaluate the success of the scoring engines. In addition, we describe a validity study that cross-validates multiple measures, including non-assessment measures. In that study we evaluate the validity of the statement that performance of the items provides evidence of the corresponding Claims in the content specifications.

Activity 6: Develop a Model for Rangefinding and for Training the Engine

Under this task we will compare in-person with presumably more cost-effective distributed, online rangefinding.

Data from the small-scale trial will be used in rangefinding and in training and validating the scoring engines. Traditional rangefinding sessions take place in person. With the geographic distribution of the SBAC states, a more distributed model will prove more cost-effective. Materials from rangefinding will feed into the scoring process, forming the foundation for human scoring. Depending on the final automated scoring approach taken and the nature of the items, the results of the rangefinding may feed directly into the scoring engines.

Activity 7: Materials and Processes for Rangefinding and Engine Training

Using student responses from the small-scale trials, AIR will develop a process for selecting responses for rangefinding and for training the engine. As discussed under Activity 6, we will test two different models for rangefinding-traditional, in person meetings and distributed rangefinding conducted over the Internet.

Activity 8: Conduct Rangefinding

AIR will conduct rangefinding by using the response samples selected under Activity 7, implementing the models described under Activity 6. Each rangefinding session, on line or in person, will be attended by an AIR content expert and a DRC scoring expert. AIR will work with SBAC leadership and state representatives to identify potential participants for the rangefinding committees.

Activity 9: Conduct Hand Scoring

Under Part 3 of this proposal we describe the details of our approach to hand scoring the responses from the trials. Because these scores will be used to train and validate the engines, each response will be double scored and every discrepancy will be resolved. Under Part 3 we also describe how we will use this opportunity to test different approaches to scoring items and tasks.

Activity 10: Train Engine and Evaluate Success, Validate Against Human Scoring for Different Types of Prompts

Under this task, AIR will train the engines and repeat the analysis described under Activities 2 and 3. Using data from the small-scale trials on items and tasks written to SBAC specifications, we will train the engine on a subsample of the available responses. We will evaluate the success of these trials by using the same analyses developed under Activity 3. These analyses will be conducted separately for different types of constructed-response items. These analyses will identify the scoring approaches that are successful for each type of item.

Activity 11: Reporting

AIR will prepare a report documenting the activities under this part of the contract. We envision the report to be iterative, with AIR delivering a draft to SBAC, revising the report based on SBAC comments, delivering the revised report to the Technical Advisory Committee, and making final revisions in response to their comments.

The final report will address two main topics: Rangefinding and scoring approaches. In discussing rangefinding, the report will document the models of rangefinding tested during this phase of the project, and summarize findings and recommendations. The report will document the scoring engines developed and tested, and provide recommendation about the types of items and tasks to which each may be applicable. The report will describe how automated scoring might be augmented with human scoring to increase the validity of the scores.

A draft outline of this report appears in the project approach.

Task 3: Item/Task/Stimulus Research and Development

Task 3 of this project has two components: a series of cognitive lab studies and a series of small scale trials, which will administer approximately 3,200 items to 150-200 students each. These activities must occur in time to inform large-scale item development.

The greatest risk to this project is timeline. It is critical that results from this research inform item development for the pilot test, and that leaves very little time for the research itself. Because the pilot test items will form the anchor set for the field-test and subsequent SBAC scale, it is also critical that the research be accurate.

Two key features of our approach will enable us to meet these critical objectives on the designated timeline:

- We have proposed to deliver the items in the cognitive labs and the small scale trials (which will include over 30,000 students) using AIR's well-proven, existing, operational test delivery system, enabling us to deploy the items very quickly and ensuring reliable delivery.
- We have proposed a phased research plan timed to deliver results to a phased item development plan at the critical juncture where that information will be required. A schematic illustration and overview of this phasing is presented in our Project Approach.

In order to ensure that items can be reliably delivered during the pilot test itself, support SBAC development of similar items in the future, and ensure the use of our system at this juncture does not inhibit competition, we are willing to open-source our item renderers for item types used in this study. The renderers allow the delivery of items, manage the student interactions, and capture student responses.

Our phased research plan begins with the cognitive labs, which will investigate the largest, riskiest innovations, and return results immediately for use in the development of pilot-test items and inform the study design for the small-scale trials.

Item development for the pilot test will defer aspects of development until key results are available. The small-scale trials confirm (or contradict) findings from the cognitive labs, and inform decisions about media used in stimuli, presentation of items in different formats, and the use of different layouts. These aspects are readily deferred where necessary until findings from the trials are available.

In our project approach we describe the work sequentially, and the tasks enumerated in this work plan follow that organization. These plans are structured and sequenced differently than those enumerated in the chart on page 29 of the RFP, so Table 21 provides a mapping, showing the tasks in which the RFP-identified activities will be conducted.

Table 21: Mapping Between RFP Requirements and Work Plan Tasks

Task identified in RFP	Work accomplished under Proposal Activity
Develop and review multiple prototypes for new/innovative stimulus type and item/task type by content area and grade level.	Activity 4
Conduct standardized cognitive labs for item/task types approximately 10 students per item/task including interviews of students after completing tasks.	Activity 2
Develop cognitive lab and small-scale trial administration materials (e.g., paper and online policies and procedures).	Activity 2
Conduct small-scale trials of item/task types to inform potential revision of item/task/stimulus specifications, established materials and procedures. The vendor should not assume the availability of school-level resources for this activity, however if any school-level support is needed to conduct these small-scale trials, the vendor will work with SBAC leadership and member states to minimize disruptions to school environments.	Activity 10
Develop sampling plan, in collaboration with the Psychometric Services contractor, for small scale trials including about 150-200 students per item/task.	Activity 10
Score items/tasks from small scale trials to test scoring rules and automated scoring.	Activity 12
Involve schools that have been recruited for participation.	Activity 10
Obtain teacher input in a variety of modes on a range of items/tasks, stimuli, and administration materials by content area and grade level.	Activity 2
Use cognitive labs and small-scale trials as an iterative development process, such that recommended revisions are implemented/evaluated/validated as improvements are identified.	Activity 2, Activity 4, Activity 9, Activity 10
Use trials and cognitive labs as an opportunity to explore access issues.	Activity 2
Produce reports containing clear recommendations for any aspect of item/task/stimulus materials and development processes based on the results from all research sources including: o Cognitive labs with students o Small scale trials with students o Teacher input on items/tasks, stimuli, and administration materials	Activity 7, Activity 15
Conduct cognitive labs for a minimum of 480 (combined total) selected response, technology enhanced and constructed response items, and 40 performance tasks across grade level bands and content areas. Emphasis should be placed into technology enhanced and constructed response items and performance tasks. These items will be specifically produced for cognitive lab purposes, and are not considered part of the total specified in Part 1.	Activity 2
Conduct a minimum of two small-scale trials for a minimum of 3200 (combined total) selected response, technology enhanced and constructed response items, and 64 performance tasks across grade level bands and content areas. Emphasis should be placed into technology enhanced and constructed response. These items will be specifically produced for cognitive lab purposes, and are not considered part of the total specified in Part 1.	Activity 10

The details of our plans to accomplish the work are presented in the Project Approach. The work plan presented here summarizes the key activities through which the objectives will be met.

Activity 1: Develop Cognitive Lab Questionnaire, Training Materials, Sampling Plan, and Focus Group materials

AIR staff will draft:

- Cognitive lab questionnaires;
- Training materials for cognitive interviewers;
- Focus group materials; and
- Sampling plans.

We propose to conduct the cognitive lab interviews in areas where we and our partners have offices, reducing the time required to establish the necessary logistics. This distributed approach will ensure diversity among the cognitive labs subjects, but will also require the training of many staff members from the diverse partner organizations.

AIR will draft the think-aloud protocols and student questionnaires, and develop standardized training materials to use with all cognitive interviewers.

We intend to gather feedback from teachers on items, questionnaires and interview plans through a series of focus groups, and we will also draft the materials (described in our project approach) to support these focus group studies.

Finally, while the cognitive labs will not be based on a scientific sample, we will design a plan to achieve diversity on a variety of demographic and geographic characteristics, including student needs for accommodation. AIR will draft this plan, and work with the psychometrics contractor to refine it.

Activity 2: Recruit Students for Cognitive Labs, Recruit Teachers for Focus Groups, Train Cognitive Lab Administrators, and hold Focus Groups

In order to recruit as broad a sample as possible, we will advertise the cognitive laboratories through multiple venues. While the most convenient approach involves recruiting through schools, the speed with which the studies must be conducted may preclude or limit this possibility. In our experience, it often takes three to six months for a school district's central office to approve recruitment of students for a project such as this and the timeline of this project will not allow that much time before we begin the recruitment process. We will work with the SBAC states where we intend to do interviewing to see if there is any way to get this permission more quickly.

The success of this phase of the project does not require recruitment through schools. In recent years we have had success advertising cognitive laboratory opportunities on Craig's List, so we will post a notice on the local Craig's Lists for the areas where we are doing interviewing. However, not everyone has access to the Internet, so we do not want to limit our recruiting to people who answer an online advertisement. We will also develop flyers advertising the cognitive laboratories and distribute those flyers at places where students are likely to congregate, including Boys and Girls Clubs, athletic facilities, libraries, community centers, and places of worship.

Activity 3: Final Report of Focus Groups

The cognitive labs will be an iterative process, with findings memoranda issued quickly when any findings hold implication for subsequent interviews or immediate findings for item development. Information from the cognitive labs will feed quickly back into the item development process. The results of the cognitive labs will feed into the item development for the small-scale pilots, and findings will inform the item development for the pilot tests.

After all the cognitive labs have been completed, AIR will produce a report on the cognitive labs that at a minimum includes the following:

- Number and type of questions administered;
- Demographic characteristics of the cognitive lab participants;

- Key findings related to specific item types;
- Efforts made to modify items based upon these findings;
- Results of interviews conducted after modifications were made; and
- Research questions identified during the cognitive lab interviews that should be further explored during the small scale trials.

Activity 4: Item and Performance Task Development for Cognitive Labs and Small Scale Tryouts

Our team will develop items for the cognitive labs and small scale tryouts using a thorough internal and external review process. The item development will proceed on a rolling basis. The outcomes from focus groups, cognitive labs and small scale tryouts will drive item development. Each of these research events will yield information about the items which in turn will contribute to item validity. AIR proposes to use focus groups during the initial phase of development in order to obtain educator input on the proposed innovative item formats. We envision the results of these focus groups as helping to shape the issues that should be addressed in the cognitive labs.

Our team will use CTB's item development tool CIAS as a database that maintains each version of the item with all of the associated stimulus and item specific attributes. In some cases, where item types are needed that are not supported by CIAS, we may supplement this tool with others. CIAS will also facilitate the linear review process that AIR will use when developing items. The reviews conducted on each item will include a series of AIR internal reviews as well as SBAC member reviews including Bias/Sensitivity.

The bulk of the items to be developed for the cognitive labs will be items measuring deeper-learning constructs that have been hard to measure in the past, as well as items that use technology in ways that have not already been heavily studied.

The items for the small scale trials will be designed as an experiment, systematically varying known characteristics of items to ensure that the ensuing data analysis will provide clear, unambiguous guidance for moving forward.

Our project approach offers the details of these plans.

Activity 5: Run Cognitive Labs

We anticipate that we will recruit approximately 145 students per grade to participate in the cognitive laboratory interviews with the goal of having 10 students respond to each of the items. Approximately 60 students per grade will respond to a single performance task (3 reading and 3 mathematics performance tasks per grade) and the remainder of the students will respond to a mix of eight other items. Each cognitive laboratory interview will take 60 to 90 minutes, which in our experience is an ideal length of time for getting the maximum information from a student without tiring the student so that he or she is unable to meaningfully focus on the task. Students will be paid $60 ($50 for their participation plus an additional $10 to cover any transportation expenses they may have).

Our plans for training interviewers and conducting the cognitive interviews with students and focus groups with teachers are described in our project approach. We propose to conduct the studies in diverse locations. The partners participating in this proposal have offices in a variety of states and to keep the costs of the cognitive labs reasonable, we anticipate that the majority of the cognitive lab interviews will be conducted in states where we have offices. These may include:

- California, the District of Columbia, Maryland, North Carolina, Oregon, Illinois, and Hawaii: AIR
- California: CTB
- Minnesota: DRC
- New York: College Board

We recognize that these are not all SBAC states, but the rapid timelines suggest that it is better to have the cognitive lab activities up and running as quickly as possible. AIR and our partners are prepared to conduct the studies in geographically diverse areas, and will emphasize SBAC states where possible.

Activity 6: Data Analysis of Cognitive Labs

Cognitive lab results are intended to be analyzed qualitatively rather than quantitatively. The cognitive lab interviews can help us identify types of questions or aspects of questions that may be causing students difficulties that are unrelated to the construct being measured. The cognitive lab think aloud protocol and the debriefing questions asked after a student completes each task will help us to identify these types of issues. Each interviewer will be given a standard template to complete at the end of each interview so the interviewer can summarize the results of the interview. Interviewers will be instructed to complete the template immediately after the interview while the student responses are still fresh in the interview's mind and to submit the completed templates to the project management team for review twice a week.

AIR research scientists will review the summaries, and where necessary, the videos and identify regularities. We will explicitly evaluate qualitative evidence of the validity of the measurement claims, factors influencing usability and access, and other questions addressed by the study.

Activity 7: Technical Report of Cognitive Labs

After all the cognitive labs have been completed, AIR will produce a report on the cognitive labs that at a minimum includes the following:

- Number and type of questions administered;
- Demographic characteristics of the cognitive lab participants;
- Key findings related to specific item types;
- Efforts made to modify items based upon these findings;
- Results of interviews conducted after modifications were made; and
- Research questions identified during the cognitive lab interviews that should be further explored during the small scale trials.

Activity 8: Develop sampling plan, test block design, and teacher questionnaires for Small Scale Tryouts

Our project approach describes our approach to developing these plans and questionnaires. Because the studies are necessarily dependent, it is not possible to offer final plans until a) the final item specifications become available; and b) results of the cognitive labs begin to flow. However, we define our approach in some detail.

We plan to develop items and tasks that vary in known ways, and assemble them into forms that will yield comparable samples for comparisons that will address key research questions. For example, for items we propose to develop items in sets of six, with items that measure the same content in different ways in each set. Each set would contribute two items to each form on which the set appeared.

This design provides for a very powerful sample. Clustering and positive intra-class correlation will help reduce the sampling variance of the key comparisons, whereas this situation often reduces the statistical power in a study. The details are described in the project approach.

Using this flexible approach our team will be able to quickly specify research questions and get the study into production.

Activity 9: Create Small Scale Trial Test forms

Our project plan describes our approach for assembling items into forms for the study. It also describes our plan for distribution of performance tasks.

Activity 10: Run Small Scale Trials

Our project approach describes our approach to recruiting schools to participating in the small scale trails, and conducting these trials using AIR's test delivery system.

AIR will draft a letter to be distributed to each of the schools selected for the small-scale trials explaining the purpose of the trial and what each school is being requested to do. AIR will consult with SBAC to determine if the letter should be sent from SBAC, from the state superintendent in each state, or from AIR. AIR will include a toll-free number in the letter that schools can call if they have any questions. AIR will include a form with the letter for schools to fill out, indicating their willingness to participate in the small-scale trials and also designating a contact person (probably the school test coordinator) for the trials. AIR will follow up with all schools that do not return the letter. In cases where we are unable to persuade a selected school to participate in the trials we may ask SBAC or a representative from the state's assessment office to follow up with the school. However, AIR will make several efforts to contact the school and gain their participation before asking for help from SBAC or the state.

While the RFP specifies that we should not rely on the availability of school resources to administer tests, in our experience indicates that virtually all schools have computer labs that they can make available for limited testing. So, while we will not rely on the availability of these resources, we will use them where available.

One key advantage of AIR's test delivery system is that it requires only the installation of a secure build of the Mozilla browser on the student computers. No expertise, servers, or other resources are required at the school.

Our sampling plan will prefer schools that can accommodate this. However, if it is necessary to deliver loaner laptops to some of the schools for use during the study, our team is prepared to do that.

Activity 11: Range Finding for Small Scale Trials

Rangefinding has two key purposes in the small scale trials:

- Obtaining materials necessary to constructed response items. These scores will be used in the validation of measurement claims about scores assigned by scoring engines and in the validation of measurement claims in general; to
- To test two different models of rangefinding.

Our project approach describes two models of rangefinding—one an adaptation of traditional rangefinding (with interactive items projected on a screen), and the other using a distributed model over webex.

These processes are described in more detail under Part 2 of this proposal.

Activity 12: Hand Scoring of Small Scale Trials and Machine Scoring

Handscoring for this project will be led by DRC. Recall that the responses will be used to train and validate scoring engines. Therefore, it is critical to have the most valid scores possible. We propose to double-score all prompts and resolve all non-matching responses. These details appear in our Project Approach. There, we describe our training, quality assurance systems, available quality assurance reports, and other plans.

As part of the scoring endeavor, SBAC would like to evaluate different types of rubrics. We have budgeted for a second rubric for 10 percent of the responses, and will include these scores in the handscoring. When SBAC specifications are available we can identify the particular types of rubrics to be evaluated.

Activity 13: Data Analysis of Small Scale Trials and Machine Scoring Validation

Our project approach describes our proposed analysis of data from the small-scale trails. The statistical approach is very similar to the approach used in Part 2 of this proposal to validate the scoring engines.

Recall that the small scale trials have three objectives:

- To provide student responses on which to train, evaluate, and refine scoring engines;
- To test or confirm the findings from the cognitive labs in larger, more diverse samples;
- To systematically evaluate the impact of technology enhancements on the validity and fairness of the tests.

The final objective, analysis of scores and their use in training, evaluating, and refining the scoring engines are described under Part 2.

We propose to address the other two questions through a mix of simple analyses and more sophisticated structural equation analysis. The structural equation analyses will all build on the same framework. That framework is a second-order latent trait model, in which items load on the claims for which they are intended to serve as evidence, and those claims load on a common subject-area trait.

These sophisticated analyses will target systematic differences across students' performances associated with different presentation and interaction factors. Much simpler analyses are appropriate to see if some approaches generally remove (or construct) barriers to performance. We propose to compare scores on items of each type to scores achieved by overlapping sets of students on items of other types designed to measure the same or similar content. Systematic differences across modes, presentations, or other features would suggest that the features associated with the less difficult item variants universally reduce barriers relative to the other modes.

To evaluate fairness across identifiable groups we will compare the number of items with each feature that are flagged for differential item function. For this study we propose to use the standard Mantel-Haenszel procedure (Holland, 1985; Holland & Thayer, 1988) for dichotomous items and the standard mean difference method (SMD) will be used for polytomous items. Using flagging criteria negotiated with SBAC, we will investigate whether items with some modes, formats, presentation or other features are more or less likely to be flagged for differential item function.

One specific question in which SBAC is explicitly interested is in the effect of mixing item types and media on a single form. As described in the illustrative item development plan above, we plan to construct forms within each grade and subject to support this study. Some forms will be homogenous, including limited and controlled item and media types. Simple comparisons across common populations tested with the different types of forms will evaluate whether student performance appears affected by mixing item types on the same form.

The analyses described here are illustrative. Until the final specifications are done and the hypotheses honed by cognitive labs, we expect that these plans will continue to evolve.

Activity 14: Technical Report of Small Scale Trials

The final report will describe the design, protocols, data, analysis and conclusions from these studies. It will follow a format similar to the Part 2 final report, but will address the issued raised in the analysis section here.

Task 4: Study of Item Procurement Options

The independent evaluation of the three SBAC item procurement options (SBAC-managed, State-managed, and State-submitted) will be planned and implemented in close collaboration with SBAC and potential participating states. The study will be implemented by the Collaborative and independently evaluated by HumRRO. We anticipate the following activities as part of this study.

Work Plan Activity 1: Define Implementation Parameters

The Collaborative will work closely with key SBAC staff to delineate and refine the definitions of each procurement option and the parameters for state participation in each. For the state managed option, we will establish guidelines for recruiting, training, and monitoring item writers, and for editing, reviewing, and revising items. For the state submitted option, guidelines for aligning existing items will be established.

Work Plan Activity 2: Design Sampling Plan

The Collaborative will work closely with SBAC to design the sampling plan for the item procurement study, including evaluation of recommendations for identifying participating states across the options.

We will work with SBAC and the governing states to identify the numbers of states and items to be included in the study.

Work Plan Activity 3: Establish Implementation Timeline

The Collaborative will work with SBAC to refine the implementation timeline to include items from each option in the 2012-13 Pilot Test so that pilot test data, along with results from independent item reviews, can be used when evaluating items developed under each option. The outcome of this work task will be to obtain approval for the timeline for the following milestones:

- Project start
- States recruited for/assigned to the three procurement options
- Items developed/submitted
- Independent item reviews conducted
- Evaluation analyses and reporting

Work Plan Activity 4: Define Evaluation Methodology

The evaluation methodology will include procedures to document the item development processes implemented under each option and identify outcome measures of cost and quality. We will work closely to define the data collection protocols and procedures for working with states within SBAC overall governance structures to define state participation parameters.

Work Plan Activity 5: Implement Procurement Options

Once the evaluation methodology has been developed and approved by SBAC, the procurement strategies will be implemented and data will be collected and evaluated. As part of the Collaborative, HumRRO will work closely with participating states to define and control the study parameters.

Work Plan Activity 6: Conduct Independent Item Reviews

Once items have been developed through the state-managed option or identified through the state-submitted option, HumRRO will conduct independent item reviews as described above for all three options including a review of items from the SBAC-managed option.

Work Plan Activity 7: Complete Evaluation Analyses and Prepare Final Report.

After all data has been collected, HumRRO will complete all analyses of the three procurement options. Once evaluation activities have been completed, a final report will be prepared. The evaluation report will describe evaluation procedures, outcomes, and recommendations that highlight strengths to continue as well as improvements to enhance the process.

C. Project Schedule

An overall program schedule has been developed to ensure the successful delivery of work products and deliverables (Appendix I). The proposed schedule includes the sequencing of events within and across tasks to ensure that the pool of stimuli, items and performance tasks for the Pilot Test meet the requirements of the Item/Task Specifications to fully measure the standards and skills intended by the

CCSS. Table 22 shows the high level pacing of the planned work tasks for each of the four tasks described in the RFP.

Table 22: Task 1, Item/Task/Stimulus Development and Reviews

Work Plan Activity	Schedule	Participants
Activity 1: Development Planning Meeting	Early April	SBAC, Collaborative
Activity 2: Stimulus Development	Early April	Collaborative
Activity 3: Stimulus Review	April-May	SBAC, Collaborative
Activity 4: Identification of Item Authors	April-May	Collaborative
Activity 5: Item Development Training	May-June	Collaborative
Activity 6: Item Development	June - September	Collaborative
Activity 7: Editing of items for Committee Review	June-September	Collaborative
Activity 8: Reviewer Training	August	Collaborative
Activity 9: Item Reviews	Sept. - Oct.	SBAC, Collaborative
Activity 10: Final review with SBAC	Oct. - Nov.	SBAC, Collaborative
Activity 11: Final Revisions to Items	November	SBAC, Collaborative
Activity 12: Final Deliverables Upload	December	Collaborative

Table 23: Task 2, Automated Scoring and Scoring Tables

Work Plan Activity	Schedule	Participants
Activity 1: Gather and Prepare Existing Student Responses	Award-April	SBAC, AIR, CTB
Activity 2: Gather Open-Source Components	Award-April	AIR
Activity 3: Conduct Initial Analysis	May-June	AIR
Activity 4: Prepare Vision Document	April	SBAC, AIR
Activity 5: Prepare Recommendations for Validation of Scoring Models	April-May	AIR
Activity 6: Develop a Model for Rangefinding and for Training the Engine	April-May	AIR, DRC
Activity 7: Materials and Processes for Rangefinding and Engine Training	Oct-Nov	AIR, DRC
Activity 8: Conduct Rangefinding	Nov-Dec	AIR, DRC
Activity 9: Conduct Hand Scoring	Nov-Dec	AIR, DRC
Activity 10: Train Engine and Evaluate Success, Validate Against Human Scoring for Different Types of Prompts	Dec-Mar	AIR
Activity 11: Prepare Final Reports And Specifications	April 2013	AIR

Table 24: Task 3, Item/Task/Stimulus Research and Development

Work Plan Activity	Schedule	Participants
Activity 1: Develop Cognitive Lab Questionnaire, Training Materials, Sampling Plan, and Focus Group materials	March-April	SBAC, AIR
Activity 2: Recruit Students for Cognitive Labs, Recruit Teachers for Focus Groups, Train Cognitive Lab Administrators, and hold Focus Groups	March-April	SBAC, AIR, CTB, DRC
Activity 3: Final Report of Focus Groups	April-May	AIR
Activity 4: Item and Performance Task Development for Cognitive Labs and Small Scale Tryouts	March – August	AIR, CTB, DRC, CAE
Activity 5: Run Cognitive Labs	April – June	AIR, CTB, DRC
Activity 6: Data Analysis of Cognitive Labs	May-June	AIR, CTB, DRC
Activity 7: Technical Report of Cognitive Labs	July	AIR
Activity 8: Develop sampling plan, test block design, and teacher questionnaires for Small Scale Tryouts	May-June	AIR
Activity 9: Create Small Scale Trial Test forms	July-August	AIR
Activity 10: Run Small Scale Trials	September – October	AIR
Activity 11: Range Finding for Small Scale Trials	October- November	AIR, DRC
Activity 12: Hand Scoring of Small Scale Trials and Machine Scoring	November 2012 – March 2013	AIR, DRC
Activity 13: Data Analysis of Small Scale Trials and Machine Scoring Validation	March 2013-April 2013	AIR
Activity 14: Technical Report of Small Scale Trials (Validation of Scoring Engine appears under Part 2)	April 2013	AIR

Table 25: Task 4, Study of State Procurement Options

Work Plan Activity	Schedule	Participants
Activity 1: Define Implementation Parameters	April 2012	SBAC, CTB, HumRRO
Activity 2: Design Sampling Plan	April 2012	SBAC, CTB, HumRRO
Activity 3: Establish Implementation Timeline	May 2012	SBAC, CTB, HumRRO
Activity 4: Define Evaluation Methodology	May 2012	SBAC, CTB, HumRRO
Activity 5: Implement Procurement Options	May-October 2012	SBAC, CTB, HumRRO
Activity 6: Conduct Independent Item Reviews	December 2012	HumRRO
Activity 7: Carry Out Evaluation Analyses and Prepare Final Report	March 2013 - June 2013	HumRRO

Our program team will use the schedule to carefully manage and monitor the program to ensure all customer deliverables and contract requirements are met within the agreed upon timeline. The Program Schedule Analyst (PSA) will be responsible for updating information, to ensure SBAC has access to reporting on all scheduled task. Each assigned program team member has specific roles and responsibilities related to the schedule to make certain that successful implementation of these tasks occur. The PSA and program management team will work closely with the SBAC team to continuously review key tasks and dates, provide status, and make any adjustments to the timelines as required and agreed upon. The PSA will also outline the critical path with the program team and customer, to highlight all critical dates and ensure key milestones are met. Throughout all phases of the program, the PSA continuously monitors and analyzes the schedule, looking downstream to ensure all dates on the critical path are on target leading to key milestones. Working closely with the Project Managers, the schedule analyst helps to ensure that the detailed departmental schedules remain in alignment with the program schedule. If any impacts to the schedule are identified, the PSA will immediately notify the Program Manager and work with the team to bring the schedule back into alignment with the customer deliverable requirements.

D. Deliverables: Fully describe all deliverables to be submitted under the proposed contract.

Task 1: Oversight, Item/Task/Stimulus Development and Reviews

Task 1 deliverables include 1A Oversight, 1B (L1-12), 1C (L1-18), 1D (L1-19), and 1E (L1-20) deliverables. Additional deliverables from the Oversight section will provide overall documentation of the development process across all four RFP tasks and will be discussed separately. Oversight reports for Task 1 will detail comments on the various item development materials developed by other vendors and will include notes on their implementation and any suggested revisions or additions.

- Item/task/stimulus development - evaluation of training materials (e.g., workshop materials, evaluations, checklists). This report will include notes on the implementation of these materials and recommendations for enhancements or additions to the materials to improve the efficacy of author training.
- Content, bias and sensitivity, and accessibility reviews - evaluation of training materials (e.g., workshop materials, evaluations, checklists). Feedback from facilitators and participants across these three review's training sessions will be compiled into a single report with recommendations for enhancements or additions to each set of training materials (see Task L1-18 and L1-19 deliverables discussion below).
- Content, bias and sensitivity, and accessibility reviews - evaluation of procedures and materials (e.g. agendas, documentation procedures, reviewer checklists, evaluations). This report will provide recommendations based on data collected from the review meeting facilitators and participants. (see Task 1C and 1D deliverables discussion below).

The primary deliverable from Task 1B will be a final set of approximately 10,000 items and performance tasks including the following components:

- Final items/tasks for the 2012-2013 Pilots as described in SBAC RFP 14, Table 1, page 34
- Stimulus materials associated with items/tasks including permission and copyright information
- All associated metadata as describe in the Item/Task Specifications from SBAC RFP 04
- Review histories and other required process documentation

These tasks will be provided in the agreed-upon format, including conformance to required interoperability standards, and appropriate for upload into the SBAC Item/Task Authoring and Pooling System.

During the process of developing the stimuli, items and tasks under Task 1, several intermediate deliverables will be completed and submitted to SBAC for approval. These intermediate deliverables include:

- Stimulus/Item development plan. This plan will include the distribution of stimuli and items to be developed across the blueprint for each content area and grade level and will be based on an analysis of the item specifications and blueprints developed under SBAC 09.
- Information on stimulus and item/task developers. While individual contracts will not be presented to SBAC for approval, CTB will provide data on the developers selected for review in accordance with the participation policies developed under SBAC 08.
- Specific plans for stimulus/item/task facilitator and developer training. The overall training plan has been described above in Section B. A detailed schedule of trainings designed in accordance with SBAC participation policies will be developed for SBAC approval. SBAC staff, work group members, and other identified representatives will be invited to attend any or all of the online training or meetings.
- Permissions/copyrights. In addition to individual information that is part of the item metadata, CTB will provide additional documentation of permissions and copyrights including limitations and costs.

Tasks 1B and 1C encompass the content, bias/sensitivity, and accessibility review meetings. These activities contribute to the development of the final item deliverable described above. During these activities, several additional deliverables will be created.

- Documentation of reviewer feedback. Reviewer feedback will be collected in accordance with review protocols described in the SBAC 08 policies and stored for later use by SBAC during final item review and reconciliation.
- Implementation notes and recommendations for improvements in training materials. These notes will be documented for inclusion in the oversight report described above.

Task 1E encompasses the final review and approval of all Pilot Test items. In support of the final item deliverable of the item pool, additional documentation will be provided to SBAC.

- Documentation of all decisions relating to final item edits and revisions according to SBAC protocols.
- Evaluation and recommendations for improvement in or enhancements to the item/content specifications. These specific recommendations will be included in the Oversight reports described above.

Task 2: Automated Scoring and Scoring Models

Table 26 presents key deliverables and approximate deliverable dates for the individual deliverables arising from these activities.

Table 26: Timeline for Key Deliverables

Task	Deliverable	Date
Task 1: Gather and Prepare Existing Student Responses	Letter to states requesting participating	April 1
	Transcription rules document	April 4
	Data file containing transcribed student responses	May 30
Task 2: Gather Open-Source Components	Design documents for 4 engines based on open source components (these include technical specifications)	April 30

Task	Deliverable	Date
	Demonstration of scoring engines	July 30
	Open-source licenses for scoring engines	July 30
	Final delivery of scoring engine code	September 30
Task 3: Conduct Initial Analysis	Report on initial analysis	September 30
Task 4: Prepare Vision Document	Draft vision document	September 30
	Final vision document	<2 weeks after task 3 end>
Task 5: Prepare Recommendations for Validation of Scoring Models	Validity Study Report	December 15
Task 6: Develop a Model for Rangefinding and for Training the Engine	Report describing rangefinding models	September 30
Task 7: Materials and Processes for Rangefinding and Engine Training	Materials to support rangefinding	October
Task 8: Conduct Rangefinding	Rangefinding meeting	November
	Final rubrics, exemplars, and selected validity and training papers	December
Task 9: Conduct Hand Scoring	Daily throughput, reliability, and validity reports	November-December
	Final data set	December
Task 10: Train Engine and Evaluate Success, Validate Against Human Scoring for Different Types of Prompts	Final dataset with auto-assigned and validated human scores	March 2013
	Final report	March 2013

Task 3: Item/Task/Stimulus Research and Development

The following deliverables are anticipated as part of the work on Task 3 relating to cognitive labs and small-scale trials.

Focus Groups:

- Materials
- Sampling Plan for Teacher Selection
- Run Focus Groups
- Final Report of Focus Group Findings

Item Development:

- Required numbers of items per item format developed per grade level per content area
- Required numbers of Performance Tasks developed per grade level per content area

Cognitive Labs:

- Student Cognitive Lab Protocols
- Sampling Plan for Student Selection

- Training Materials
- Run Cognitive Labs
- Technical Report of Cognitive Labs

Small Scale Trials:

- Sampling Plan of Schools and Students for Small Scale Trials
- Recruitment Plan for Small Scale Trials
- Data Analysis Plan for Small Scale Trials
- Block Design of Test Forms for Small Scale Trials
- Test Forms for Grade Levels and Content Areas
- Conduct Small Scale Trials
- Conduct Range Finding
- Conduct Hand Scoring
- Conduct Machine Scoring
- Data Analysis of Small Scale Trials and Validation of Scoring Engine
- Technical Report of Small Scale Trials

Task 4: Study of Item Procurement Options

CTB and HumRRO have defined the following deliverables for Task 4.

- Sampling Plan: Plan to include the number of states participating in each procurement option, description of how states were selected/assigned to the options, stratification of states to options, number of items per item type needed by states in each option, and description of how items will be selected for administration.
- Evaluation Plan: Description of research plan to complete a systematic and fair comparison of the three procurement options, including procedures to document item development processes implemented under each option; list of cost and quality outcome measures; data to be collected and sources; and analyses to be completed.
- Evaluation Report: Description of review processes (e.g., content, bias/sensitivity, accessibility) and findings; assessment of item quality for each procurement option; assessment of costs, timelines, materials, and feasibility for each procurement option; evaluation of relative success of each procurement option; and recommendations for implementing option with greatest likelihood of sustainability.

E. Outcomes and Performance Measurement: Describe the impacts/outcomes the Vendor proposes to achieve as a result of the delivery of these services including how these outcomes would be monitored, measured and reported to SBAC.

The Collaborative is committed to the Consortia goal of providing a balanced assessment system to monitor student growth along achievement continua so that states can monitor students who are on track to reach college and career readiness in high school. Since the scores from the summative and interim assessments are to be used to 1) establish both the status and progress of student learning, 2) progress toward college and career readiness, and 3) evaluate the performance of schools and teachers, the final pool of stimuli, items, and tasks must provide sufficient evidence of student learning to support these uses.

In addition to the specific outcomes described below for each task, the following qualitative questions should also be considered relative to the overall success of the project. While these are longer term outcomes related to the success of the SBAC balanced assessment program, making them difficult to measure during the course of this contract, they represent critical issues that should guide and be the focus of the development of these deliverables.

Will the stimuli, items and tasks developed for the Pilot Test result in an assessment system that is valid, reliable and fair to students?

Will the stimuli, items and tasks developed for the Pilot Test result in assessments that elicit relevant evidence to assign students to achievement levels?

Will the development processes and materials developed for the Pilot Test inform the development of both a summative and an interim assessment system that will be useful to educators?

Task 1: Oversight, Item/Task/Stimulus Development and Reviews

In addition to the individual task outcomes defined by SBAC, we have defined three overarching measures of success for the project. The complete set of deliverables, taken together, should ensure that these outcomes are accomplished.

1. The pool of stimuli, items and tasks for the 2012-2013 Pilot Test will be reflective of the SBAC Content Specifications and Item/Task Specifications and will elicit sufficient evidence to support claims about student learning. Stimuli, items and tasks will arise from the appropriate application of the SBAC framework of Content Specifications, Item and Task specifications, Test Specifications, and psychometric designs and will represent the instantiation of these specifications and designs within the body of the Common Core State Standards.
2. The stimuli, items and tasks developed for the Pilot Test pool will be informed by research during the development process. Data from the development and testing of automated scoring models and algorithms as well as the cognitive labs and small-scale trials, which focus on technology-enhanced items, will inform the authoring of items. The final pool of items will incorporate the important findings from these research activities.
3. The technical quality of the deliverables will be sufficient to meet the test purposes defined in the purpose statements of RFP 09. All stimuli, items, and tasks will have been reviewed sufficiently to ensure their adherence to the item specifications, bias/sensitivity guidelines, and accessibility and accommodations guidelines and will result in valid and reliable assessments.

Oversight- Outcomes

While CTB will manage all parts of this project, other Collaborative members will contribute to achieving identified outcome for this section of Task 1. We have defined the following outcomes and performance measures related to the oversight activities describe in Task 1 in Table 27.

Table 27. Oversight Outcomes and Performance Measures

Outcome	Description/Performance Measure
Effective collaboration with SBAC	Close collaboration with SBAC and our partners is critical to the success of the project. We envision that our initial planning meeting will include discussions on SBAC expectations for preferred processes for communication, identification of key contact persons, decision makers, and roles so that a communication plan can be established for the project..
Effective management within the Collaborative	We are proposing a senior program management team with vast experience in managing collaborative partnerships to meet successful outcomes and delivery through use of PMI methodologies and industry best practices. Clear set expectations, requirements, defined metrics in collaboration with SBAC for approval of each deliverable, and communication with all stakeholders will help drive on time and high quality delivery. These partnerships coupled with

Outcome	Description/Performance Measure
	effective program management will allow us to provide SBAC with leading industry experience and expertise to meet project expectations.
Maintenance of all timelines and deliverables	As part of our program management team, a program schedule analyst has been assigned to provide continual oversight and monitoring of key tasks that lead to final deliverable timelines. The schedule analyst works hand in hand with the project team and PM to track project status through key "real time" and scheduled project benchmark checkpoints. The Program Manager will report to project schedule status to SBAC through regular reporting mechanisms including weekly management reports and quarterly and annual project reports. Any risk to the final deliverable schedule will be mitigated by addressing internal timelines so that SBAC deliverable timelines are not impacted.
Consolidation of work across the project	Program Management will work closely with the Project Director to coordinate the work plans across the four tasks of the project with particular attention to ensuring that feedback from the research activities impacts the development process. Success for this outcome will be measured by the final alignment to the specifications for automated scoring and technology enhancements and ultimately by data from the Pilot Tests.
Communication and record-keeping	Program Management will maintain key project documentation including Meeting Agendas, Records and Minutes, Weekly Status Reports, Records of Decision Making and, Action logs, and Quarterly and Annual Project Progress Reports, CTB will propose a format for all project communication/record keeping, but is amenable to adopting formats currently in use with current vendors based upon SBAC preferred formats.
Fiscal and organizational responsibility for meetings	Program Management will maintain fiscal and organizational responsibility for all project meetings. Logistics information will be available through established collaboration channels. Key fiscal and meeting summary information will be provided through weekly management reports, Quarterly and Annual Project Progress Reports.
Reports of recommendations for improvement	Recognizing that the activities under this contract will be using the SBAC specifications and training materials for the first time, the Collaborative will carefully document recommendations and share with SBAC throughout the implementation process and use of these materials in order to provide real time feedback and recommendations for consideration to allow continuous improvement.. Our team will also conduct a thorough post project review with all stakeholders to create a final project report that will include recommendations for future enhancements to SBAC materials, as well as, vendor process improvements recommended to allow us to better serve your needs in the future. This documentation will be organized into recommendations for improvements or enhancements to the materials and provided to SBAC. The recommendations will cover all aspects of the processes and procedures that are guided by SBAC documentation.

Being strong proponents of the belief that we can only "manage what we can measure", key performance metrics will be identified, tracked and reported through regular project team "stand-up" meetings instituted to promote collaboration, ensure the team is fully aware of project status, and can remain focused on critical tasks and deliverables. The project team will be supported by project status dashboards and metric reporting tools that report key performance indicators (displayed through collaboration software) in a manner that supports the successful management, tracking and reporting project status. Stand-up meetings have been part of CTB's management culture for years, and have been used very effectively with our business partners in the SBAC proposal development process. The team building and ability to focus on critical issues quickly and effectively make this an important tool in our proposed team management plan.

Task 1: Item/Task/Stimulus and Reviews

We recognize that the development of stimuli, items, and tasks for the 2012-2013 Pilot Test will be the first time that the Item/Task Specifications, Test Specifications, Stimulus, Bias/Sensitivity and Accessibility Guidelines, Item Writer and Reviewer Training Materials and other documents have been used to develop items for a large-scale administration of an SBAC assessment. We anticipate close coordination with SBAC at all levels of development to ensure that the specifications and training materials are providing sufficient basis for development and to provide mid-course adjustments as needed. We will carefully adhere to the communication plans described in the management plan and are committed to both formal and informal means of communication with SBAC as work under this contract proceeds. All decisions and the implementation will be documented to inform further item development under other contracts. The primary outcome of this collaborative effort will be a set of implementation notes and recommendations for enhancements to the above materials.

Item/Task/Stimulus Development and Reviews Outcomes

We describe specific performance measures below for each significant outcome and as an additional evaluative measure we propose a set of more qualitative questions about the eventual impact of the items beyond their use in the Pilot Test. For most outcomes related to stimulus, item and performance task development, the performance measures will consist of an evaluation of the deliverables against the SBAC Content Specifications, Item/Task Specifications, Test Specifications, and other guidelines that define the work under this RFP. See Table 28 for a summary of these outcomes and performance measures.

Table 28: Content Development and Reviews

Outcome	Description/Performance Measure
Comprehensive content development plan	This initial outcome will provide a plan for the SBAC-managed development of stimuli, items, and tasks that cover the range of content (claims and assessment targets) including item type and cognitive complexity needed for the 20122-2013 Pilot Test. The Collaborative will work closely with SBAC staff and work groups to develop a comprehensive content development plan. The plan will be evaluated against the requirements of the item specifications and blueprints. Once approved, regular updates on the actual development with respect to the content development plan will be provided to SBAC to allow for any needed changes to the plan, reallocation of resources, or other input from SBAC representatives.
Wide range of stimulus materials	This outcome is closely linked to the content development plan. The development of diverse stimuli will be guided by the criteria in the specifications. The plan for obtaining permissioned stimuli location and developing new stimuli will be part of the overall content development plan. The final set of stimuli will align closely to the SBAC Stimulus Specifications.
Management and tracking of permissions and copyrights	All details related to the permissioning and copyrights for stimuli will be tracked and managed. This outcome can be measured by the comprehensive documentation provided for permission and copyright tracking.
Inclusion of innovative stimuli, items/tasks that cover the full range of SBAC specifications and provide evidence of student learning beyond traditional measures	The Collaborative will provide a diverse collection of stimuli and items/tasks from a combination of teacher and Collaborative development efforts. CTB, AIR, DRC, the College Board and CAE will all contribute to the development of content to meet the Item and task specifications, providing a wide range of items and tasks in response to the specifications.

Outcome	Description/Performance Measure
Attention to complete scoring descriptions, including automated scoring considerations	This outcome will be achieved by including sufficient training and specifications for item writers around the development of scoring rules and rubrics for conventional items and for the inclusion of automated scoring considerations for constructed-response and technology-enhanced items as defined in the Item and Task specifications and in guidance received from the automated scoring research carried out under Task 2.
Inclusion of diverse types of technology-enhanced items	While the field is still developing clear definitions of technology-enhanced items, the Collaborative will work closely with the appropriate SBAC work group to refine the preliminary definitions of technology-enhanced items in the RFP to include a wide range of technology enhancements. This outcome can be evaluated by the adherence of the items produced to the technology-enabled and technology-enhanced item specifications.
Recruiting and training of item/task writers and stimulus, item and task reviewers	The collaborative will manage all details of recruitment, contracting, and payment of individuals who participate in the authoring and review of items/tasks and stimuli. We will work closely with SBAC to ensure that the pool of individuals identified for this work matches the SBAC participation Policies and other guidance from SBAC and its member states.
Development of a pool of stimulus/items/tasks free from bias	Specific attention will be given to the incorporation of Bias/Sensitivity guidelines during item authoring, and review committees will be recruited to include individuals with specific expertise in reviewing stimuli and items for bias and sensitive issues. The comments provided by reviewers during the Bias/Sensitivity review will serve as a measure of the effectiveness of the direction given by the Bias/Sensitivity Guidelines and the appropriate sections of the training materials. These reviews and the documentation of comments will be carried out in accordance with SBAC policies and protocols.
Development of a pool of stimulus/items/tasks that are accessible to all students	Specific attention will be given to the incorporation of Accessibility and Accommodations guidelines during item authoring, and review committees will be recruited to include individuals with specific expertise in reviewing stimuli and items for accessibility. The comments provided by reviewers during the Accessibility and Accommodations review will serve as a measure of the effectiveness of the direction given by the Accessibility Guidelines and the appropriate sections of the training materials. These reviews and the documentation of comments will be carried out in accordance with SBAC policies and protocols.
Documentation of review feedback	The Collaborative will carefully documental all reviewer feedback during the stimulus and item/task reviews. This feedback will be captured electronically and available for SBAC review both informally (through participation in the reviews) or formally during the review and reconciliation process. This feedback will not only provide a basis for revisions to the items and tasks, but can also provide general feedback on the efficacy of the training and development processes.
Reconciliation of feedback	While the Collaborative expects to include SBAC representatives in all steps of the development process, reconciliation of review comments will be a targeted activity between the Collaborative management and SBAC staff, work group members, or other identified representatives. The collaborative will recommend revisions based on SBAC protocols and policies, and provide sufficient documentation and time for SBAC to review all recommendations and make final revision decisions.

Outcome	Description/Performance Measure
Maintenance of metadata and review histories	Al metadata and post-review revisions will be tracked during the item development process for both teacher-authored items and Collaborative–developed items. The metadata fields will adhere to the tagging and other requirements laid out in the stimulus/item/task specifications and any additional requirements for interoperability from the SBAC Item Authoring and Pooling System. The final upload and quality review of content in this system will provide an evaluation of the alignment between provided metadata with the stated requirements.
Develop final revised stimuli/items/tasks in close collaboration with SBAC	The delivery of a pool of items that meets all specifications is the overall outcome of this contract. The successful upload and quality review of the poll of items for the Pilot Test in 2012-2013 will be the culmination of the various outcomes described above.

Alignment Study

As an additional performance measure to document the alignment of the final item pool to the item and task specifications, CTB proposes to carry out an alignment study of the final item pool. We will collect alignment data from the review checklists in the authoring system after the content reviews and analyze to determine how the final item pool aligns to the requirements of the standards with regard to two factors: alignment to the content standard and alignment to the cognitive complexity required by the specifications. The results of this study will serve as a measure of the effectiveness of the processes and materials used to develop the item pool.

Task 2: Automated Scoring and Scoring Models

The key outcome from the set of activities described under Part 2 of this contract is a specification describing a comprehensive approach to scoring. If successful the approach will use automated scoring as a central component and will provide scores that can be validly used as evidence of the content claims that the items are designed to support.

Below, we enumerate interim and final outcomes, along with performance metrics.

1. On-time delivery of each deliverable.
2. Delivery of working scoring engines that meet or surpass the state of the art on existing prompts prior to data available from the small scale trials. Success in this endeavor can be evaluated using the following metrics:
3. Ability to match or exceed rates of inter-human rater agreement
4. Ability to match or exceed the match of existing, proprietary scoring engines to final resolved scores from humans
5. Ability of the engine to self-identify responses with a high-risk of incorrect scoring. This measure will include a tradeoff between false positives and false negatives, and the optimal mix will be determined in conjunction with SBAC.
6. Delivery of technical documents that can be used successfully for code-review of the open-source engines by independent SBAC experts. These technical document will comprise the technical specifications.
7. Delivery of final scoring engines that meet or surpass the state of the art on responses SBAC prompts collected in the small-scale trials. Success in this endeavor can be evaluated using the following metrics:
8. Ability to match or exceed rates of inter-human rater agreement
9. Ability to match or exceed the match of existing, proprietary scoring engines to final resolved scores from humans

10. Ability of the engine to self-identify responses with a high-risk of incorrect scoring. This measure will include a tradeoff between false positives and false negatives, and the optimal mix will be determined in conjunction with SBAC.
11. Delivery of a vision document that, by independent review, reflects an optimal mix of cost savings and validity within the capabilities of the state-of-the art. This project may extend the state of the art.

Task 3: Item/Task/Stimulus Research and Development

Answers to research questions comprise the primary outcome of these activities. These objectives focus on new and innovative stimulus materials and item/task types across content areas and grade levels.

We identify both interim and overall performance measures. Key outcomes and measures of success in this endeavor include:

- Early identification of specific item types, task types, measurement approaches and stimulus types requiring investigation;
- On-time conduct of the time-critical cognitive labs
- The design of the small scale trials will explicitly identify the hypotheses to be tested
- A power analysis of the small scale trial design will show it able, within the scope of the contract, to detect meaningful effects with 80 percent power.
- The trials achieve the response rate assumed in the power analysis.
- Scoring will be completed with reliability of final scores to be negotiated with SBAC.

Task 4: Study of Item Procurement Options

The overall outcome of the study will be a systematic comparison of the three procurement strategies. Comparison information will be conveyed to SBAC via a report that (a) documents the procedures followed to develop items/tasks/stimulus materials and (b) describes the relative success of each option. This information should enable SBAC to determine the procurement option that is most feasible and cost effective approach to implement and that results in producing high quality items/tasks/stimulus materials.

F. Risks: Define risks you identify as being significant to the success of the project. Include how you would propose to effectively monitor and manage these risks, including reporting of risks to the SBAC's contract manager.

Each of the Collaborative partners strives to focus on process to ensure that there are effective controls in place to consistently deliver customer requirements. We strive to implement process improvements to enhance customer satisfaction and reduce costs to the customer.

To help support these efforts, the CTB program and project teams are trained in and implement effective risk management techniques focused on risk management planning, risk identification, analysis, responses, and monitoring and controlling risks on a project. Risk management is critical because it equates to preventing problems, which is fundamental to increased customer satisfaction and increased efficiency and quality.

CTB conducts a risk assessment prior to beginning work. The goal is to identify potential non-conformances in advance and address them by putting controls in place to prevent them from occurring. Our standard procedure is to do this in conjunction with the customer to ensure we define and understand the "Critical to Satisfaction" requirements of a project and to listen to the voice of the customer. In preparation for the initial program meeting, the program manager will prepare a standard risk analysis for the SBAC Pilot Items/Tasks/Stimulus Research, Development and Reviews program.

Through initial discussions and follow up with SBAC, risks and controls planned to prevent them will be clarified and agreed upon with SBAC leadership. The risk plan will be managed in conjunction with SBAC as the program work progresses.

The program to develop stimuli, items and tasks for the 2012-2013 Pilot Test has risks that are typical of complex programs, with some additional challenges that require attention and control. General program risks follow, and specific risks associated with each task are listed later.

In addition to the risks posed by an aggressive schedule and the complexity of tasks, there are some risks associated with the overall assessment system of which the Pilot Test item pool will be a part.

- Fidelity between the SBAC Content Specifications, Item/Task Specifications and other SBAC guidelines and the final pool of Pilot Test items for SBAC's balanced assessment system and goals of providing practically useful information for teaching and learning and delivering cost-effectiveness for sustainability.
- We are proposing a focused review at the inception of the program to mitigate the risk of conceptual misunderstanding between our interpretation of these documents and SBAC expectations. We anticipate that there will be close collaboration between the Collaborative staff and SBAC work group members throughout the item development process. SBAC staff can, as is feasible, participate in all or part of our proposed training, development, and review sessions.
- Lack of coherence between item and task specifications, training materials, test designs, blueprints, proficiency level descriptors and other documents due to timing constraints and the completion of these assessment components by various vendors.
- The Technical Director, Sally Valenzuela and key development staff will work closely with SBAC as we implement these materials for the first time. We will monitor how the training materials, specifications documents, and guidelines are being used by teachers and reviewers throughout the process, and will communicate regularly with SBAC about modification to the materials or processes that might be made on an ongoing basis. Any changes to the materials during implementation will be discussed and approve by SBAC as well as described in the formal deliverable documentation of suggested changes to these documents.
- Productive and creative interaction between the Collaborative and the members of SBAC's leadership and work group structure that enables the work and reviews to be completed thoughtfully in the available time.
- Initial attention on creating effective lines of communication and establishing management tools will mitigate the risk of lost information and opportunity across the SBAC organizational structure and the Collaborative. We will provide communication structures to work closely with SBAC to ensure that our management solutions are effective and remain that way through the completion of the program.
- Efficient collaboration and decision making that facilitates optimal use of time
- Our expert resources will become familiar with the item/task specifications, training materials, guidelines, and other SBAC documents as soon as they are available. The moment the program begins, we will be prepared to present draft plans for the deliverables and the implementation plans can be adjusted based on our deliberations with SBAC. Our preparations to achieve the agreement and approval of SBAC to proceed with work will enable the time available to be fully utilized.

In addition to these overall risks, we have identified specific risks for each of the four tasks described in the RFP. The following tables describe these task-specific risks and our proposed mitigation strategies. At the outset of the contract, we will meet with SBAC to discuss these risks and refine the mitigation and backup plans.

Task 1: Oversight, Item/Task/Stimulus Development and Reviews

Table 29. Task 1 Risks and Mitigation Strategies

Risk	Mitigation Strategy
Communication issues resulting from interdependent tasks with overlapping timelines	Reporting hierarchy across partner organizations established, with necessary commitment from organizations, to ensure absolute clarity in project authority/responsibility. Collaborative structure implemented to ensure necessary project interfaces - includes daily Stand-Up, status reporting dashboards - aided by collaboration tools. Clarity in status reporting/escalation paths. Comprehensive project documentation - Work Plan, Master Schedules, Deliverable Matrix - in place to provide transparency in project status, cadence, key events. Staff assignments - assignment of Senior level Program Manager with necessary body of experience and proven success in managing complex program across contributing organizations.
Insufficient time to locate/develop stimuli prior to beginning of the item/task authoring window	Collaborative partners will begin mobilizing resources to write/locate stimuli before contract award. Our experienced staff will search for the 45% of stimuli that are previously published, as they are familiar with rights holders that permit use of their material in assessments. CTB's Permissions Department is also skilled at tracing copyright holders to investigate whether materials presumed to be in the public domain truly are.
Insufficient numbers of SBAC educators recruited for item authoring	CTB will work closely with the SBAC08 contractor and leadership and member states to contract with sufficient numbers of teachers for item authors, include overages to allow for attrition. As a backup plan, we will increase the numbers of items/tasks developed directly by the Collaborative using professional item writers.
SBAC educators are less experienced in content, pedagogy, and assessment than anticipated	Our model for small groups facilitated by an editor allows for ongoing training and feedback to educators as the author items. ; sufficient numbers of authors will be recruited to allow for some attrition of teachers either voluntarily or excused according to participation policies.
Schedule challenges prevent research input from cognitive labs, small-scale trials and automated scoring research to be available at the beginning of item authoring	Our model for distributed item writing and editing, in addition to the three-phase approach, allows multiple entry points for research findings; more challenging technology-enhanced items can be written later in the process. As a backup plan, additional review time can be added to incorporate specific review/revision rounds related to automated scoring guidelines and important research findings.
SBAC educators produce insufficient numbers of items for committee review	CTB will provide ongoing tracking of items authored and their editing status within the item authoring system and will provide regular updates on progress toward the fulfilling the content development plan. The Collaborative can adjust the number of vendor-produced items as a backup plan.
Insufficient numbers of reviewers for bias, sensitivity and accessibility reviews.	The Collaborative will contract with a sufficient overage of reviewers to ensure all reviews can be completed on schedule. Our model of online reviews over a period of several weeks will provide flexibility that should allow adequate reviewer participation.

Risk	Mitigation Strategy
Time constraints for SBAC members to review the final item pool	The Collaborative will define all review windows well in advance to provide for flexibility in scheduling. We will use transparent systems and processes that will allow SBAC monitoring and, if desired, participation at all levels of the process. We will explicitly facilitate appropriate SBAC participation throughout the authoring process and reviews so that the final deliverables will meet SBAC expectations and will not require time-consuming reviews and revisions.
Technical issues with final upload requiring item edits after upload (ex. missing metadata)	The Collaborative will work closely with the SBAC 07 vendor to identify and meet all requirements for the final upload. Our work plan includes quality reviews during the vendor editing round to ensure items match specifications and contain all required data elements. We propose a trial upload prior to the approval of the final deliverable to identify and address potential issues.

Because of the response to Q105 in the Q&A, we have not addressed the risks of late deliverables from other vendors but are we are prepared to discuss contingency plans with SBAC should this occur.

Task 2: Automated Scoring and Scoring Models

As noted above, the timeline risks to the activities of Task 2 are critical. The small-scale trials constitute a key milestone on the critical path to develop and validate the scoring engines. To mitigate the risk of waiting for results from the small scale tryouts, our team proposes begin the evaluation of automated scoring solutions almost immediately upon contract award. To support this schedule, we will begin with existing student responses, which we will solicit from SBAC member states.

The lack of quality open-source solutions is a risk that we address by identifying in this proposal the best candidate solutions. We have also mitigated risk by assigning staff with extensive experience with both natural language processing and online test delivery. Backup plans include the evaluation of existing proprietary engines, along with plans to ensure competition if those solutions become necessary.

Table 30. Task 2 Risks and Mitigation Strategies

Risk	Mitigation Strategy/Backup Plan
Late delivery of research findings to development process	Begin evaluation immediately upon contract award with existing student responses
Lack of open-source scoring solutions that meet SBAC requirements	Experienced Collaborative resources to evaluate; evaluate proprietary solutions if required

Task 3: Stimulus/Item/Task Research and Development

Table 31. Task 3 Risks and Mitigation Strategies

Risk	Mitigation Strategy	Backup Plan
Item specifications not available on time	The phased nature of this plan mitigates this risk somewhat, this event is on the critical path	SBAC will have two options: Move forward testing draft specifications or delay the timeline
Slow approvals of critical path plans and materials	Circulation of early draft documents and clear plans up front. Key staff will work individually with SBAC representatives to obtain approval	SBAC will have two options: move forward on provisional plans or delay the timeline

Risk	Mitigation Strategy	Backup Plan
Too few cognitive lab participants recruited	Multiple recruitment paths and a rolling recruitment/interview schedule	Reduce the number of items tested or delay timelines
Too few schools recruited for small scale trials	Multiple recruitment paths, multiple contacts, an recourse to SBAC assistance	Reduce the scope or power of the small scale trials or delay the timeline

Task 4: Study of Item Procurement Options

Table 32. Task 1 Risks and Mitigation Strategies

Risk	Mitigation Strategy
Limited Sample of Participating States	Expand efforts to recruit more states to participate; include cautions regarding when reporting findings and offering recommendations
States Unable to Produce/Submit Requested Items	Begin communications with participating states immediately upon contract award and maintain regular correspondence; provide support to states, as appropriate, to secure item writers, reviewers, etc.